

معالجة اللغات الطبيعية للويب الدلالي

دراسات ١٦

د. ديانا ماينارد
د. كالينا بونتشيفا
د. إيزابيل أوغنتشتاين

ترجمة

د. خالد بن عبدالرحمن الميمان

دراسات ١٦

معالجة اللغات الطبيعية للويب الدلالي

تأليف

د. ديانا ماينارد
جامعة شيفلد
المملكة المتحدة

د. كالينا بونتشيفا
جامعة شيفلد
المملكة المتحدة

د. إيزابيل أوغنتشتاين
جامعة كلية لندن
المملكة المتحدة

ترجمة

د. خالد بن عبدالرحمن الميمان
جامعة القصيم
المملكة العربية السعودية

١٤٤٠هـ - ٢٠١٩م





معالجة اللغات الطبيعية للويب الدلالي

الطبعة الأولى
١٤٤٠ هـ - ٢٠١٩ م

المملكة العربية السعودية - الرياض
ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa

مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية، ١٤٤٠هـ.
فهرسة مكتبة الملك فهد الوطنية أثناء النشر
ماينارد، ديانا
معالجات اللغات الطبيعية للويب الدلالي./ ديانا ماينارد؛ كالينا بونتشيفا؛ ايزابيل اوغنتشتاين ؛ خالد بن عبدالرحمن الميمان. - الرياض، ١٤٤٠هـ
..ص؛ .. سم
ردمك: ٨ -٣٥ -٨٢٢١ -٦٠٣ -٩٧٨
١ - اللغات - معالجة البيانات أ. بونتشيفا، كالينا (مؤلف مشارك) ب. اوغنتشتاين ، ايزابيل (مؤلف مشارك) ج. الميمان ، خالد بن عبدالرحمن (مترجم) د. العنوان
ديوي ٤٠٠،٢٨٥ ٦٦٧٣/ ١٤٤٠
رقم الإيداع: ٦٦٧٣/ ١٤٤٠
ردمك: ٨ -٣٥ -٨٢٢١ -٦٠٣ -٩٧٨

التصميم والإخراج
دار وجوه للنشر والتوزيع
Wojooh Publishing & Distribution House
www.wojoooh.com
المملكة العربية السعودية - الرياض
الهاتف:4562410 الفاكس:4561675
للتواصل والنشر:
info@wojoooh.com



هذه ترجمة عربية لكتاب:

Natural Language Processing for the Semantic Web
Synthesis Lectures on the Semantic Web: Theory and Technology

Published by
Morgan & Claypool Publishers (2016), United Kingdom

فهرس الكتاب

١٨١

مقدمة المترجم

الحمد لله، والصلاة والسلام على رسول الله، نبينا محمد عليه أفضل الصلاة وأتم التسليم، وبعد:

عصر الذكاء الاصطناعي كما يدعوه البعض بذلك، وأحيانا يدعى بعصر البيانات الضخمة، هذا العصر الذي تقاس فيه قوة الكيانات بما تملكه من بيانات وكيف تستطيع تحليلها والإفادة منها. يأتي هذا الكتاب في ظل شح المكتبة العربية بالمؤلفات حول هذا الفن، ويقدم للقارئ العربي المفاهيم الرئيسة لتقنيات معالجة اللغات الطبيعية، والتي تندرج تحت علم الذكاء الاصطناعي. يبسط هذا الكتاب تلك المفاهيم بداية من الكلمات وصرفها إلى تجزئة الجمل وتصنيف أقسام الكلام والتعابير الدلالية المختلفة، مرورا بأحدث التطبيقات والأدوات التي تستخدم لمعالجة اللغات الطبيعية، ثم يربط ذلك بالويب وكيف يمكن أن تتكامل تقنيات معالجة اللغات الطبيعية مع تقنيات الويب والبيانات الضخمة.

تقنيات الويب الدلالي تقوم بتحويل البيانات غير الهيكلية إلى بيانات نافعة وذات معنى، وتعد تقنيات معالجة اللغات الطبيعية من أهم وأنفع الطرق لتحويل البيانات الضخمة في الويب إلى بيانات ذات مدلول يمكن قراءتها وتحليلها والاستفادة من مخرجاتها. يندرج تحت موضوع الويب الدلالي العديد من الموضوعات المتعلقة بمعالجة

اللغات الطبيعية، ويعرض هذا الكتاب أهمها. فمن الأمثلة الحيوية التي تطرق لها هذا الكتاب موضوع تحليل المشاعر تجاه أمر ما (منتج، حدث، موقف، أو غيره). هذا الموضوع الذي تعكف على تطويره كبريات الشركات في العالم سواء التجارية منها كأمازون أو مواقع التواصل الاجتماعي مثل تويتر وغيرها في شتى المجالات التجارية والسياسية والاقتصادية والاجتماعية.

اللغة العربية تشترك مع لغات العالم كونها تتألف من جذور وجذوع وكلمات وسوابق ولواحق وجمل وحروف جر وأصوات وغيرها، وتختص مع عدد من لغات العالم كونها تكتب من اليمين لليسار، كما تختص مع عدد قليل جدا من اللغات العالمية كونها لغة ذات غنى صرفي، وتنفرد بأن الله سبحانه وتعالى شرفها بأن تكون لغة لكتابه العزيز، الذي لا يأتيه الباطل من بين يديه ولا من خلفه تنزيل من حكيم حميد.

ولذا كان من الواجب على أهل الاختصاص في اللغة العربية وأهل الاختصاص في الحاسب الآلي وهم المعنيون بالدرجة الأولى أن يعملوا جنبا إلى جنب في مجال (معالجة اللغة العربية حاسوبيا)، لتواكب بل ولتسبق اللغات الأخرى؛ فاللغة العربية تأتي في المركز الأول عالميا في عدد الدول التي أقرتها لغة رسمية فيها. وإن تكلمنا عن روعة وإتقان فصاحتها وبلاغتها فلن توفيها الكلمات حقها ولو طالت. وأشير هنا إشارة تذكير وهي أن معالجة اللغات الطبيعية لا تعني أن نطوع اللغة لتناسب مبادئ الحاسب الآلي، بل لندرب الحاسب الآلي ليفهم ويدرك اللغة ويتعامل معها كتعامل وفهم البشر قدر ما نستطيع، وهذا هو المبدأ الرئيس لعلم معالجة اللغات الطبيعية.

يقدم هذا الكتاب المفاهيم الرئيسة بشكل مبسط، ولذلك فهو من أنسب الكتب لمن يجد نفسه راغبا في الدخول إلى علم معالجة اللغات الطبيعية والويب الدلالي، حيث لايكتفي هذا الكتاب بشرح أسس علم معالجة اللغات الطبيعية وارتباطه بالويب الدلالي بل يقدم الأدوات المناسبة والحديثة المستخدمة في كل مهمة من مهام هذه العلوم، ويقارن بينها ويعرضها ببساطة، ولذا نقترح على القارئ الكريم أن تكون منهجيته في القراءة التطبيق على هذه الأدوات المقترحة أو بعضها فالتطبيق يرسخ المعلومة ويوضح اللبس فيها إن وجد.

وأشير إلى نقطة مهمة للقارئ الكريم وهي أن يلقي نظرة على مسرد المصطلحات في آخر الكتاب قبل أن يبدأ القراءة، والهدف من ذلك أن تكون كلمات المصطلحات مفهومة وواضحة ومألوفة بالنسبة له، إذ لاتوجد مصطلحات عربية موحدة في هذا المجال، ولعل هذا العمل أن يكون نقطة انطلاقة لتوحيد الجهود نحو مصطلحات موحدة ومتفق عليها من قبل المتخصصين في هذا المجال.

ولا يفوتني في هذا المقام أن أتقدم بالشكر الوافر بعد شكر الله سبحانه لهذا المركز المبارك، مركز الملك عبدالله بن عبدالعزيز لخدمة اللغة العربية، جزى الله القائمين عليه خير الجزاء ووفقهم وسددهم.

د. خالد بن عبدالرحمن الميمان

جامعة القصيم

١٥ جمادى الأولى ١٤٤٠ هـ

كلمة المحرر

سواء أكنت تسميها الويب الدلالي، أم البيانات المرتبطة، أم الويب ٣,٠، فإن الجيل الجديد من تقنيات الويب يحقق تقدمًا كبيرًا في تطور الشبكة العنكبوتية العالمية. نظرًا لأن الجيل الأول من هذه التقنية ينتقل خارج المختبرات، فإن الأبحاث الجديدة تستكشف كيف ستغير شبكة الويب المتنامية عالمنا. في حين أن موضوعات مثل بناء علم الوجود والمنطق تبقى مهمة، وهناك مجالات جديدة مثل استخدام علم الدلالة في بحث الويب، وربط واستخدام البيانات المفتوحة على الويب، والتطبيقات المستقبلية التي ستدعمها هذه التقنيات، كل هذه الاتجاهات تعد مجالات بحث مهمة.

كل مستخدمي الويب، سواء أكانوا علماء أم مهندسين أم ممارسين، يحتاجون بشكل متزايد إلى فهم أعمق -ليس فقط للتقنيات الجديدة للويب الدلالي- بل لفهم المبادئ التي تعمل بها هذه التقنيات، وأفضل الممارسات لتجميع الأنظمة التي تدمج اللغات المختلفة والموارد المتنوعة والوظائف التي ستكون مهمة في الحفاظ على شبكة الإنترنت التي تتوسع بسرعة، وتغير بشكل مستمر كميةُ المعلومات التي غيرت حياتنا.

الموضوعات المضمنة في هذا الكتاب:

- مبادئ الويب الدلالي من البيانات المرتبطة إلى تصميم الأنطولوجيا
- تقنيات وخوارزميات الويب الدلالي الرئيسة

- تقنيات البحث واللغة الدلالية
- شبكة البيانات» الناشئة واستخدامها في تطبيقات الصناعة والحكومات والتطبيقات المستخدمة في الجامعات
- الثقة والشبكات الاجتماعية وتكنولوجيا التعاون وعلاقتهم بالويب الدلالي
- اقتصاديات تكييف الويب الدلالي واستخدامه
- النشر والعلوم في الويب الدلالي
- الويب الدلالي في مجال الرعاية الصحية وعلوم الحياة

وهنا قائمة بالكتب التي تضمها سلسلة المحاضرات المجمعة حول الويب الدلالي:

Natural Language Processing for the Semantic Web
Diana Maynard, Kalina Bontcheva, and Isabelle Augenstein
2016

The Epistemology of Intelligent Semantic Web Systems
Mathieu d'Aquin and Enrico Motta
2016

Entity Resolution in the Web of Data
Vassilis Christophides, Vasilis Efthymiou, and Kostas Stefanidis
2015

Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description
Carol Jean Godby, Shenghui Wang, and Jeffrey K. Mixter
2015

Semantic Mining of Social Networks
Jie Tang and Juanzi Li
2015

Social Semantic Web Mining
Tope Omitola, Sebastián A. Ríos, and John G. Breslin
2015

Semantic Breakthrough in Drug Discovery

Bin Chen, Huijun Wang, Ying Ding, and David Wild
2014

Semantics in Mobile Sensing
Zhixian Yan and Dipanjan Chakraborty
2014

Provenance: An Introduction to PROV
Luc Moreau and Paul Groth
2013

Resource-Oriented Architecture Patterns for Webs of Data
Brian Sletten
2013

Aaron Swartz's A Programmable Web: An Unfinished Work
Aaron Swartz
2013

Incentive-Centric Semantic Web Application Engineering
Elena Simperl, Roberta Cuel, and Martin Stein
2013

Publishing and Using Cultural Heritage Linked Data on the Semantic Web
Eero Hyvönen
2012

VIVO: A Semantic Approach to Scholarly Networking and Discovery
Katy Börner, Michael Conlon, Jon Corson-Rikert, and Ying Ding
2012

Linked Data: Evolving the Web into a Global Data Space
Tom Heath and Christian Bizer
2011

المحرران:

يينغ دينغ- جامعة إنديانا

بول جروث- معامل إلسفير

الفصل الأول

مقدمة

معالجة اللغات الطبيعية (NLP) هي المعالجة التلقائية للنص المكتوب باللغات الطبيعية (البشرية) (الإنجليزية والفرنسية والصينية وغيرها)، بدلاً من اللغات الاصطناعية مثل لغات البرمجة، والغاية من تلك المعالجة هي محاولة «فهم» النص. تُعرف معالجة اللغات الطبيعية أيضًا باسم اللغويات الحاسوبية (CL) أو هندسة اللغات الطبيعية (NLE). تشمل معالجة اللغات الطبيعية مجموعة واسعة من المهام، بدءًا بالمهام ذات المستوى المنخفض، مثل تقسيم النص إلى جمل وكلمات، ووصولاً إلى تطبيقات معقدة رفيعة المستوى مثل إضافة الحواشي والشروحات الدلالية وتعدين الآراء. نقصد بالويب الدلالي إضافة الدلالات، أو المعاني، إلى البيانات الموجودة على شبكة الإنترنت، بحيث يمكن معالجة صفحات الويب والتعامل معها من قبل الآلة بسهولة كبرى أحد المظاهر الرئيسة لهذا المفهوم تتمثل في وصف الموارد باستخدام مُعرّفات فريدة، تسمى مُعرّفات الموارد الموحدة (URIs). يمكن أن تكون الموارد كيانات، مثل «باراك أوباما»، أو مفاهيم مثل «سياسي» أو علاقات تصف كيفية ارتباط الكيانات بعضها ببعض، مثل «زوجة». توفر تقنيات معالجة اللغات الطبيعية وسيلة لتعزيز بيانات الويب بالدلالات، على سبيل المثال عن طريق إضافة معلومات عن الكيانات والعلاقات بصورة تلقائية وفهم أيٍّ من الكيانات الموجودة في العالم الحقيقي تجري الإشارة إليها بحيث يمكن تخصيص مُعرّف URI لكل كيان.

الهدف من هذا الكتاب هو تعريف القراء المتعاملين مع تقنيات الويب الدلالي، أو المهتمين بها، بموضوع معالجة اللغات الطبيعية ودورها وأهميتها في مجال الويب الدلالي. على الرغم من أن مجال معالجة اللغات الطبيعية وُجِد قبل ظهور الويب الدلالي بوقت طويل، إلا أن أهميته لم تبرز على الواجهة بقوة إلا في السنوات الأخيرة، ولا سيّما مع انتقال تقنيات الويب الدلالي نحو تقنيات موجهة نحو التطبيقات بصورة كبرى. لذلك فإن الغرض من هذا الكتاب هو تفسير دور معالجة اللغات الطبيعية وإعطاء القراء فهمًا أكبر لبعض مهام معالجة اللغات الطبيعية التي تعدّ الأكثر أهمية لتطبيقات الويب الدلالي، بالإضافة إلى تقديم بعض الإرشادات حول اختيار الأساليب والأدوات الأنسب والأكثر ملاءمة لسيناريو معين. في نهاية الأمر، يتمثل الهدف في أن يخرج

القارئ بالمعرفة اللازمة لفهم المبادئ الرئيسة، وإذا لزم الأمر، المعرفة اللازمة لاختيار تقنيات معالجة اللغات الطبيعية المناسبة التي يمكن استخدامها لتعزيز تطبيقات الويب الدلالية.

سيكون الهيكل العام للكتاب كما يلي. سنصف أولاً بعض المكونات الأساسية منخفضة المستوى، ولا سيّما تلك التي توجد عادة في مجموعات أدوات العمل مفتوحة المصدر الخاصة بمعالجة اللغات الطبيعية والتي تُستخدم على نطاق واسع في أوساط المهتمين بهذا المجال. بعد ذلك سنبيّن كيف يمكن الجمع بين هذه الأدوات واستخدامها كمُدخلات للمهام ذات المستوى الأعلى، مثل استخلاص المعلومات وإضافة الحواشي والشروحات الدلالية وتحليل شبكات التواصل الاجتماعي وتعدين الآراء، وأخيرًا سنوضح كيف يمكن بناء تطبيقات على نمط التطبيقات المعززة دلاليًّا لاسترجاع المعلومات وتصورها، وتطبيقات نمذجة مجتمعات الإنترنت، على أساس تلك المهام.

هناك نقطة ينبغي أن نوضحها، وهي أنه عندما نتحدث عن معالجة اللغات الطبيعية في هذا الكتاب، فإننا نشير أساسًا إلى مهمة فهم اللغات الطبيعية (NLU) الفرعية، ولا نشير إلى مهمة توليد اللغات الطبيعية (NLG) الفرعية ذات الصلة بالمهمة السابقة. وعلى الرغم من أن توليد اللغات الطبيعية مهمة مفيدة ولها صلة أيضًا بالويب الدلالي، على سبيل المثال فيما يتعلق بتمرير نتائج تطبيقٍ ما إلى المستخدم بطريقة يمكن فهمها بسهولة، خصوصًا في الأنظمة التي تتطلب عرض النتائج بصيغة صوتية، إلا أنها خارج نطاق هذا الكتاب، لأنها تستخدم تقنيات وأدوات مختلفة جدًّا. وبالمثل، هناك عددٌ من المهام الأخرى التي لن نناقشها هنا على الرغم من كونها تندرج عادة ضمن نطاق معالجة اللغات الطبيعية، ولا سيما المهام التي تُعنى بالكلام بدلاً من النص المكتوب. ومع ذلك، تستخدم العديد من التطبيقات الخاصة بمعالجة الكلام وتوليد اللغات الطبيعية مهام معالجة اللغات الطبيعية ذات المستوى المنخفض التي سنقوم بوصفها. هناك أيضًا بعض التطبيقات رفيعة المستوى المبنية على معالجة اللغات الطبيعية التي لن نغطيها في هذا الكتاب، مثل تطبيقات التلخيص والإجابة عن الأسئلة، على الرغم من كونها تعتمد أيضًا على الأدوات نفسها ذات المستوى المنخفض.

معظم أدوات معالجة اللغات الطبيعية التي ظهرت مبكرًا، مثل المحللات النحوية (على سبيل المثال: محلل الاعتماد المفاهيمي لشانك Schank's conceptual dependency parser [1]) هذه المحللات النحوية كانت مبنية على القواعد، ويرجع ذلك جزئيًّا إلى هيمنة بعض النظريات اللغوية (نظريات نعوم تشومسكي في المقام الأول [2])، إضافة إلى عدم وجود القدرات الحاسوبية اللازمة، وهو ما جعل أساليب تعلم الآلة غير مجدية. في الثمانينيات الميلادية، بدأت أنظمة التعلم الآلي في الظهور على الواجهة، لكنها كانت تُستخدم بشكل أساسي فقط لإنشاء مجموعات من القواعد المشابهة لأنظمة القواعد المطوّرة يدويًّا كانت موجودة في السابق، وذلك باستخدام تقنيات مثل أشجار القرار. ومع اكتساب النماذج الإحصائية شعبية كبرى، خاصة في مجالات مثل الترجمة الآلية وتصنيف أقسام الكلام، حيث كانت الأنظمة المستندة إلى قواعد محكمة في كثير من الأحيان غير كافية لإزالة أوجه الغموض، وباتت نماذج ماركوف المخفية (HMMs) شائعة، مستحدثة مفهوم الخصائص الموزونة وأساليب صنع القرار الاحتمالي. وفي السنوات القليلة الماضية، اكتسب التعلم العميق والشبكات العصبية أيضًا شعبية عالية جدًّا، وذلك بعد نجاحها المذهل في مجال التعرف على الصور والرؤية الحاسوبية (على سبيل المثال في التكنولوجيا المُستخدمة في السيارات ذاتية القيادة)، على الرغم من أنه لا مجال لمقارنة ذلك النجاح الدرامي بنجاحها في مهام معالجة اللغات الطبيعية في الوقت الحالي. التعلم العميق هو في الأساس فرعٌ من فروع التعلم الآلي يستخدم مستويات هرمية متعددة من الخصائص التي يتم تعلمها بطريقة غير خاضعة للإشراف unsupervised، وهذا يجعله مناسبًا جدًّا للتعامل مع البيانات الكبيرة، لأنه يتميز بالسرعة والكفاءة، ولا يتطلب عملية الإنشاء اليدوي للبيانات التدريبية، على عكس نظم التعلم الآلي التي تتم تحت الإشراف. ومع ذلك، وكما سيتبين من خلال هذا الكتاب، فإن إحدى مشكلات معالجة اللغات الطبيعية تتمثل في أن الأدوات (البرمجية) المستخدمة تحتاج للتكيف مع نطاقات ومهام محددة في معظم الأحيان، وغالبًا ما تكون عملية تكييف الأدوات أسهل مع استخدام النظم المبنية على القواعد عندما يتعلق الأمر بمجالات التطبيق في العالم الحقيقي. وفي معظم الحالات، يجري استخدم خليط يضم أساليب مختلفة، وهذا يعتمد على المهمة المطلوبة.

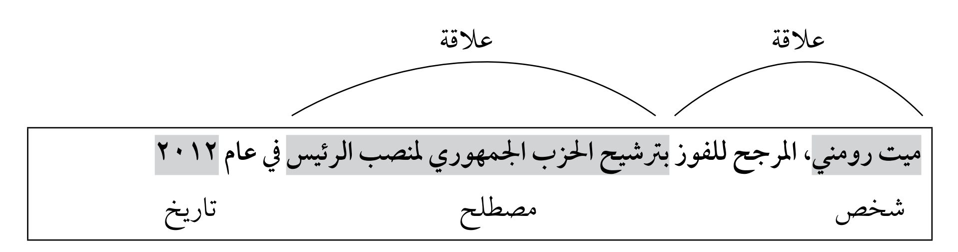
١-١ استخلاص المعلومات

استخلاص المعلومات هو عملية استخراج المعلومات وتحويلها إلى بيانات منظمة، وقد يتضمن ذلك تعبئة مصدر معرفي منظم بمعلومات من مصدر معرفي غير منظم [3]. بعد ذلك يمكن استخدام المعلومات الواردة في قاعدة المعارف المنظمة كمصدر للمهام الأخرى، مثل الإجابة على الاستفسارات التي تتم باللغات الطبيعية أو تعزيز محركات البحث العادية بأشكال معرفية أعمق أو أكثر ضمنية مقارنة بتلك المعبَّر عنها في النص. نعني بمصادر المعرفة غير المنظمة النص الحرّ، مثل النص الموجود في مقالات الصحف والمدونات وشبكات التواصل الاجتماعي وصفحات الويب الأخرى، بدلاً من الجداول وقواعد البيانات والأنطولوجيات أو التجميعات، التي تشكل نصوصًا منظمة. ما لم يُنص على خلاف ذلك، سوف نستخدم كلمة نص في بقية هذا الكتاب للإشارة إلى النص غير المنظم.

عند النظر في المعلومات الواردة في النص، هناك عدة أنواع من المعلومات يمكن أن تكون ذات أهمية. تُعد الأسماء الصحيحة من المكونات الرئيسة للنص، وتسمى أيضًا كيانات الأسماء (NEs)، وتشمل أسماء الأشخاص والمواقع والمنظمات. إلى جانب الأسماء الصحيحة، تعدُّ التعبيرات الزمنية غالبًا، مثل التواريخ والأوقات، كيانات أسماء. يبين الشكل ١-١ بعض كيانات الأسماء البسيطة في جملة. يتم ربط كيانات الأسماء معًا بواسطة العلاقات. علاوة على ذلك، يمكن أن تكون هناك علاقات بين العلاقات نفسها، على سبيل المثال العلاقة التي تشير إلى أن شخصًا ما هو الرئيس التنفيذي لشركة ما مرتبطة بالعلاقة التي تشير إلى أن شخصًا ما هو موظف في شركة ما، وذلك عن طريق علاقة خصائص فرعية، لأن الرئيس التنفيذي هو نوع من أنواع الموظفين. هناك نوع أكثر تعقيدًا من أنواع المعلومات، ألا وهو الحدث، ويمكن النظر إلى هذا النوع على أنه مجموعة من العلاقات التي ترتكز على الزمن. تتضمن الأحداث عادة المشاركين في الحدث وتاريخ البدء وتاريخ الانتهاء والموقع، على الرغم من أن بعض هذه المعلومات قد تكون ضمنية فقط. من الأمثلة على ذلك افتتاح مطعم. يوضح الشكل ١-٢ كيفية ارتباط الكيانات بالعلاقات التي تشكل أحداثًا مرتكزة على الزمن.

ميت رومني، المرجح للفوز بترشيح الحزب الجمهوري لمنصب الرئيس في عام ٢٠١٢
شخص   مصطلح   تاريخ

الشكل ١-١: أمثلة على كيانات الأسماء.



الشكل ١-٢: أمثلة على العلاقات والأحداث.

استخلاص المعلومات عملية صعبة؛ لأن هناك العديد من الطرق للتعبير عن الحقائق نفسها:

- عينت شركة BNC القابضة السيدة ج. توريتا رئيسة جديدة لمجلس إدارتها.
- خلفت جينا توريتا نيكولاس أندروز كرئيسة لشركة BNC القابضة.
- تولت السيدة جينا توريتا رئاسة شركة BNC القابضة.

علاوة على ذلك، قد تكون هناك حاجة لدمج المعلومات الموجودة في عدة جمل قد لا تكون متتالية.

- بعد نضال طويل في مجلس الإدارة، تنحى السيد أندروز من منصبه كرئيس لمجلس إدارة شركة BNC القابضة، وخلفته السيدة توريتا.

تتألف عملية استخلاص المعلومات عادة من سلسلة من المهام، وتشمل:

١. المعالجة اللغوية المسبقة (ستشرح في الفصل الثاني)؛
٢. التعرف على كيانات الأسماء (ستشرح في الفصل الثالث)؛
٣. استخلاص العلاقات و/ أو الأحداث (ستشرح في الفصل الرابع).

تمييز كيانات الأسماء (NER) هي مهمة التعرف على أن الكلمة أو سلسلة الكلمات المتعاقبة هي اسم صحيح، وغالبًا ما يتم تنفيذها بشكل مشترك مع مهمة تحديد أنواع كيانات الأسماء، مثل الشخص أو الموقع أو المنظمة، وهو ما يُعرف باسم تصنيف كيانات الأسماء (NEC). في حال تنفيذ المهام في الوقت نفسه، يشار إلى ذلك بالتعرّف على كيانات الأسماء وتصنيفها. يمكن أن يكون التعرّف على كيانات الأسماء وتصنيفها إما مهمة إضافة تعليقات وشروحات، أي إضافة ملحوظات إلى نص يحتوي على كيانات أسماء، أو يمكن أن تكون المهمة ملء قاعدة معارف بكيانات الأسماء هذه. عندما لا تكون كيانات الأسماء مجرد بنية مسطّحة، وتكون مرتبطة بكيان متناظر في أحد الكيانات المعجمية، يُعرف ذلك بالشرح التوضيحي الدلالي أو ربط كيانات الأسماء (NEL). التحشية الدلالية أقوى بكثير من التعرّف على الكيانات المُسمّاة، لأنها تتيح إجراء عمليات الاستدلال والتعميم، وذلك لأن عملية ربط المعلومات تتيح الوصول إلى المعرفة غير الواردة صراحة في النص. عندما يكون الشرح التوضيحي الدلالي جزءًا من العملية، غالبًا ما يشار إلى مهمة استخلاص المعلومات غلى أنها استخلاص المعلومات المستندة إلى علم الأنماط (OBIE) أو استخلاص المعلومات الموجه بواسطة علم الأنماط (انظر الفصل الخامس). يرتبط ذلك ارتباطًا وثيقًا بعملية تعلم الأنماط والتعبئة (OLP) كما هو موضح في الفصل السادس. تعدّ مهام استخلاص المعلومات أيضًا شرطًا أساسيًّا للعديد من مهام استخراج الآراء، ولا سيّما عندما تتطلب هذه المهام تحديد العلاقات بين الآراء وأهدافها، وحيثما تستند إلى علم الأنماط، كما هو موضح في الفصل السابع.

١-٢ الغموض

يستحيل على أجهزة الكمبيوتر تحليل اللغة بشكل صحيح ١٠٠٪، لأن اللغة شديدة الغموض. تعني اللغة الغامضة أنه يمكن تقديم أكثر من تفسير، إما من الناحية التركيبية أو الدلالية. كبشر، يمكننا في كثير من الأحيان استخدام المعرفة المتاحة في العالم لحل أوجه الغموض هذه واختيار التفسير الصحيح، لكن لا يمكن للحواسيب الاعتماد بسهولة على المعرفة المتاحة في العالم والحس السليم، لذلك تضطر لاستخدام

التقنيات الإحصائية أو غيرها من الوسائل لحل الغموض. غالبًا ما يتم تصميم بعض أنواع النصوص، مثل عنوانات الصحف والرسائل المنشورة على شبكات التواصل الاجتماعي، لتكون غامضة بشكل متعمد لغرض الترفيه أو لجعلها محفورة في الذاكرة، وفيما يلي بعض الأمثلة الكلاسيكية على ذلك:

- Foot Heads Arms Body (فوت يرأس هيئة الأسلحة).
- Hospitals Sued by 7 Foot Doctors (ملاحقة مستشفيات قضائيًّا من قبل ٧ أطباء متخصصين في القدم).
- British Left Waffles on Falkland Islands (اليسار البريطاني يراوغ بشأن جزر فوكلاند).
- Stolen Painting Found by Tree (العثور على اللوحة المسروقة بجانب شجرة).

في العنوان الأول، هناك غموض نحوي بين الاسم الصحيح (للعائلة) (Michael) Foot، والمقصود بها هنا شخص، وبين الاسم الشائع foot (قدم)، الذي يشير إلى أحد أعضاء الجسم؛ وبين كلمة heads التي قد تعني فعل (يرأس) أو اسم جمع (رؤوس)، وينطبق الأمر ذاته على الأسلحة. هناك أيضًا غموض دلالي بين معاني كلمة arms (أسلحة وأحد أعضاء الجسم)، وbody (هيكل الجسم ومجموعة كبيرة). في العنوان الثاني، هناك غموض دلالي بين معاني كلمة foot (أحد أعضاء الجسم ووحدة القياس)، وأيضًا الغموض النحوي الناتج عن طريقة ربط الصفات التعريفية (٧ [أطباء قدم] أو [٧ أقدام] أطباء). في المثال الثالث، هناك اثنان من أنواع الغموض، وهما الغموض النحوي والدلالي، في كلمة left (صيغة الماضي للكلمة، أو الاسم الجمع الذي يشير إلى السياسيين اليساريين). في المثال الرابع، هناك غموض في دور حرف الجر by (كعامل أو كموقع). في كل مثال من هذه الأمثلة، هناك معنى واحد ممكن بالنسبة للإنسان، والمعنى الآخر إما مستحيل أو مستبعد للغاية (الأطباء الذين يبلغ طولهم ٧ أقدام، على سبيل المثال). أما بالنسبة للآلة، فإن التوصل إلى فهم من دون سياق إضافي مفاده ترك معجنات الوافل [من عبارة left waffles] في جزر فوكلاند، على الرغم من كون هذا الفهم ممكنًا تمامًا، هو خبر بعيد الاحتمال، ويكاد يكون مستحيلاً.

١-٣ الأداء

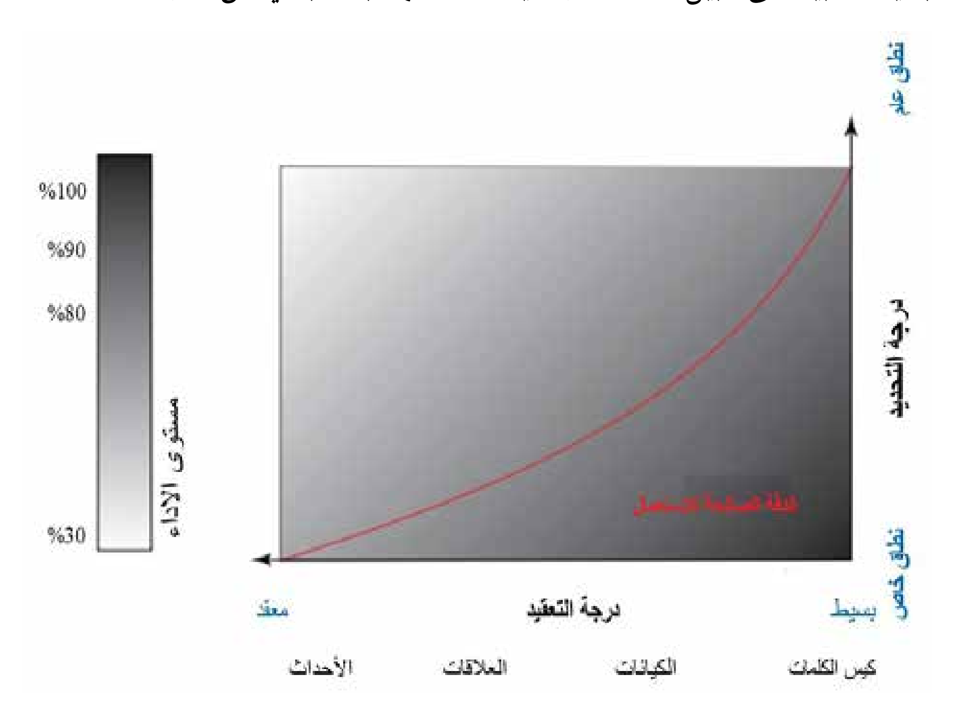
يختلف الأداء في مهام معالجة اللغات الطبيعية اختلافًا واسعًا، سواءٌ أكان بين المهام المختلفة أم بين الأدوات المختلفة، وهذا الاختلاف في الأداء ليس بسبب الغموض فحسب، بل نظرًا لمجموعة متنوعة من القضايا الأخرى، كما ستتم مناقشتها في عدة فصول من هذا الكتاب. ستتم مناقشة الأسباب التي تقف وراء اختلاف الأداء مع اختلاف الأدوات في الأقسام ذات الصلة، ولكن السبب يكمن بشكل عام في أن بعض الأدوات تكون جيدة في الأداء في بعض العناصر وفي الوقت نفسه سيئة في بعضها الآخر، وهناك أيضًا العديد من المشكلات المتعلقة بالأداء التي تبرز عندما يتم تدريب الأدوات على نوع واحد من البيانات واختبارها على نوع آخر. لكن سبب تفاوت أداء المهام على نطاق واسع يعود إلى حد بعيد إلى التعقيد.

إن تأثير الاعتماد على النطاق على فاعلية أدوات معالجة اللغات الطبيعية هي مسألة غالبًا ما يتم إغفالها. ولكن لكي تكون التقنيات مناسبة للتطبيقات في العالم الحقيقي، يجب أن تكون الأنظمة قابلة للتخصيص بسهولة لكي تناسب مجالات جديدة. تركز بعض مهام معالجة اللغات الطبيعية على وجه التحديد، مثل استخلاص المعلومات، على النطاقات الفرعية الضيقة إلى حد بعيد، كما ستتم مناقشته في الفصلين الثالث والرابع. تعرقل العديد من الاختناقات المختلفة تكيّف النظم القائمة مع مجالات جديدة، ومن هذه الاختناقات الحصول على البيانات التدريبية للنظم القائمة على التعلم الآلي. عندما يتعلق الأمر بتكييف تطبيقات الويب الدلالي، قد تكون الاختناقات في الأنطولوجيات أو التجميعات أحد الأسباب، كما ستناقش في الفصل السادس.

هناك مسألة منفصلة، وإن كانت ذات صلة، تتعلق بتكييف النظم الحالية مع أنواع مختلفة من النصوص. لا نعني بذلك التغييرات في المجال فحسب، بل أيضًا أنواع الوسائط المختلفة (مثل البريد الإلكتروني والنص المنطوق والنص المكتوب وصفحات الويب وشبكات التواصل الاجتماعي)، وأنواع النصوص المختلفة (مثل التقارير والخطابات والكتب)، والهياكل أو البنى المختلفة (مثل التخطيطات). قد يتأثر نوع النص بعدة عوامل، كالمؤلف والجمهور المستهدف ومدى كون النص رسميًّا. على

سبيل المثال، قد لا تتبع النصوص الأقل رسمية القواعد القياسية، مثل الكتابة بالأحرف الكبيرة أو علامات الترقيم أو حتى الأشكال الإملائية، وكلها عوامل يمكن أن تسبب إشكالية للآليات المعقدة لأنظمة استخلاص المعلومات. سوف تناقش هذه المسائل بالتفصيل في الفصل الثامن.

تصبح العديد من مهام معالجة اللغات الطبيعية، وخاصة المهام الأكثر تعقيدًا، عالية الدقة وقابلة للاستخدام فقط عندما تكون مركزة بشكل محكم وتقتصر على تطبيقات ومجالات معينة. يوضح الشكل ١-٣ مخططًا ثلاثي الأبعاد يظهر المقايضة بين عمومية المجال أو خصوصيته، وتعقيد المهمة، ومستوى الأداء. من هنا يمكننا أن نرى أنه يتم تحقيق أعلى مستويات الأداء في مهام معالجة اللغة التي تركز على مجال محدد والتي تكون بسيطة نسبيًّا (على سبيل المثال: تحديد كيانات الأسماء أبسط بكثير من تحديد الأحداث).



الشكل ١-٣: المفاضلات في مستويات أداء مهام معالجة اللغات الطبيعية.

لكي تكون عملية دمج تطبيقات الويب الدلالي مجدية، يجب أن يكون هناك نوع من التجانس المنطقي المقبول بين العاملين في حقل الويب الدلالي وحقل معالجة اللغات

الطبيعية. ينطبق هذا الأمر بالطبع على جميع التطبيقات التي تتطلب دمج معالجة اللغات الطبيعية. على سبيل المثال، يحتمل أن تكون بعض التطبيقات التي تندرج تحت موضوع معالجة اللغات الطبيعية غير قابلة للاستخدام فعليًّا في العالم الحقيقي كنظم تلقائية مستقلة قائمة بذاتها دون تدخل بشري. لكن الأمر ليس كذلك بالضرورة عندما يتعلق الأمر بأنواع أخرى من تطبيقات الويب الدلالي التي لا تعتمد على معالجة اللغات الطبيعية. بعض التطبيقات مصممة لغرض مساعدة المستخدم البشري بدلاً من أداء المهمة بشكل مستقل تمامًا. كثيرًا ما تكون هناك مفاضلة أو مقايضة بين مقدار الاستقلالية التي ستعود بأعلى قدر من المنفعة على المستخدم النهائي. على سبيل المثال، تمكّن نظم استخلاص المعلومات المستخدم النهائي من تفادي قراءة مئات أو حتى آلاف الوثائق بالتفصيل من أجل الحصول على المعلومات التي يريدها، لأن البحث في ملايين الوثائق يدويًّا يكاد يكون من المستحيل. من ناحية أخرى، يجب على المستخدم أن يضع في اعتباره أن أي نظام يعمل بشكل آلي بالكامل لن يكون دقيقًا بنسبة ١٠٠٪، وأنه من المهم أن يكون تصميم النظام مرنًا من حيث المفاضلة بين دقة المعلومات والقدرة على استرجاعها. بالنسبة لبعض التطبيقات، قد يكون من المهم استرجاع كل شيء، على الرغم من أن بعض المعلومات التي يتم استرجاعها قد تكون غير صحيحة. من ناحية أخرى، قد يكون من المهم أن يكون كل شيء يتم استرجاعه دقيقًا، حتى لو فُقدت بعض الأشياء.

١-٤ هيكل الكتاب

تم تصميم كل فصل من فصول الكتاب بهدف عرض مفهوم جديد في مسارات مهام معالجة اللغات الطبيعية، وشرح كيف يبنى كل مكوّن بالاعتماد على المكوّنات السابقة التي جرى وصفها. في كل فصل، نشرح المفهوم العام للعنصر، ونقدم أمثلة على الأساليب والأدوات الشائعة. وعلى الرغم من أن كل فصل يعدُّ مستقلاً بذاته إلى حد ما، من حيث كونه يشير إلى مهمة محددة، إلا أن الفصول يبنى بعضها على بعض، ولذا فإن أفضل طريقة لقراءة الفصول الخمسة الأولى لهذا الكتاب هي بالتتابع.

يصف الفصل الثاني المنهجيات الرئيسة المستخدمة في مهام معالجة اللغات الطبيعية، ويشرح مفهوم مسارات مهام معالجة اللغات الطبيعية. بعد ذلك يتم وصف مكونات المعالجة اللغوية التي تتكون منها مسارات المهام- بما في ذلك التعرف على اللغة وتجزئة الجمل وتقسيم الجمل وتصنيف أقسام الكلام والتحليل الصرفي والتحليل اللغوي والتقطيع - وتُقدم أمثلة على بعض مجموعات أدوات معالجة اللغات الطبيعية الرئيسة.

يقدم الفصل الثالث مهمة التعرف على كيانات الأسماء وتصنيفها (NERC)، وهي عنصر أساسي في استخلاص المعلومات ونظم إضافة التعليقات والشروحات الدلالية، كما يناقش الفصل أهميتها وقيودها، إضافة إلى تلخيص المنهجيات الرئيسة لهذه المهمة، ووصف مسارات المهام النموذجية المُستخدمة في مهمة التعرف على كيانات الأسماء وتصنيفها.

يشرح الفصل الرابع مهمة استخلاص العلاقات القائمة بين الكيانات، ويوضح كيف ولماذا يكون ذلك مفيدًا لعملية التعبئة التلقائية لقواعد المعارف. يمكن أن تندرج المهمة إما على استخلاص العلاقات الثنائية بين كيانات الأسماء، أو استخلاص علاقات أكثر تعقيدًا، مثل الأحداث. كما يشرح هذا الفصل مجموعة متنوعة من المنهجيات ومسارات مهام استخلاص العلاقات النموذجي، ويعرض التفاعل بين مهمة التعرف على كيانات الأسماء ومهمة استخلاص العلاقة، إلى جانب مناقشة التحديات البحثية الرئيسة.

يوضح الفصل الخامس كيفية القيام بعملية ربط الكيانات عبر إضافة الدلالات إلى أحد نظم استخلاص المعلومات القياسية غير المهيكلة من النوع الذي تم وصفه في الفصول السابقة. يناقش هذا الفصل سبب كون عملية استخلاص المعلومات غير المهيكلة غير كافية لكثير من المهام التي تتطلب وفرة أكبر في المعلومات ومزيدًا من الاستنتاجات المنطقية، ويوضح كيفية ربط الكيانات التي تم العثور عليها بأحد الكيانات المعجمية وموارد البيانات المفتوحة المترابطة مثل DBpedia وFreebase. كما يقدم الفصل أمثلة على مسارات المهام المُستخدمة عادة في إضافة التعليقات والشروحات الدلالية، وكذلك أمثلة على التطبيقات في العالم الحقيقي.

يقدم الفصل السادس مفهوم التطوير الآلي للكيانات المعجمية أو الأنطولوجيات اعتمادًا على نص غير منظم، حيث يتضمن هذا المفهوم ثلاثة مكونات مترابطة هي: التعلم والتعبئة والتنقيح. كما تتم مناقشة بعض هذه المصطلحات وكيفية تفاعلها، والعلاقة بين تطوير الكيانات المعجمية والتحشية الدلالية، ويتم وصف بعض المنهجيات النموذجية، ويتم البناء مرة أخرى على المفاهيم التي سبق عرضها في الفصول السابقة.

يشرح الفصل السابع طرق وأدوات الكشف عن أنواع مختلفة من الآراء والمشاعر والعواطف وتصنيفها، ويظهر مرة أخرى كيف يمكن تطبيق عمليات معالجة اللغات الطبيعية التي سبق شرحها في الفصول السابقة على هذه المهمة. على وجه الخصوص، يمكن أن يستفيد تحليل المشاعر المستند إلى الخصائص (مثل العناصر المحبوبة أو المكروهة في منتج ما) من عملية دمج الكيانات المعجمية الخاصة بالمنتجات في المعالجة. كما يتم تقديم أمثلة على تطبيقات حقيقية في مختلف المجالات، وهو ما يبين كيف يمكن أيضًا إدخال تحليل المشاعر في تطبيقات أوسع في عمليات تحليل شبكات التواصل الاجتماعي. ونظرًا لأن تحليل المشاعر غالبًا ما يتم تطبيقه على شبكات التواصل الاجتماعي، يُفضل قراءة هذا الفصل بالاقتران مع الفصل الثامن.

يناقش الفصل الثامن المشكلات الرئيسة التي تتم مواجهتها أثناء تطبيق تقنيات معالجة اللغات الطبيعية التقليدية على نصوص شبكات التواصل الاجتماعي، نظرًا لاستخدامها غير العادي وغير المتسق لقواعد الإملاء والنحو وعلامات الترقيم وغيرها من الأمور. ولأن الأدوات التقليدية لا تقدم أداءً جيدًا في كثير من الأحيان عند تعاملها مع هذه النصوص، فإنها غالبًا ما تتطلب أن يتم تكييفها مع هذا النوع من النصوص. على وجه الخصوص، يمكن أن تترك المكونات الأساسية للمعالجة الأولية التي سبق شرحها في الفصلين الثاني والثالث تأثيرًا خطيرًا على العناصر الأخرى الموجودة في مسارات المهام إذا ما ظهرت أخطاء في هذه المراحل المبكرة. يقدم هذا الفصل بعض الطرق الحديثة لمعالجة نصوص شبكات التواصل الاجتماعي ويعطي أمثلة على بعض التطبيقات الحقيقية.

يجمع الفصل التاسع بين جميع العناصر التي ورد شرحها في الفصول السابقة من خلال تعريف ووصف عدد من مجالات التطبيق التي تتطلب إضافة تعليقات وشروحات دلالية، مثل استرجاع المعلومات وتصورها بطريقة معززة دلاليًّا، وبناء نماذج المستخدمين الدلالية الاجتماعية، ونمذجة مجتمعات الإنترنت. كما يتم وصف المنهجيات وأدوات المصدر المفتوح الشائعة في هذه المجالات، بما في ذلك التقييم وقابلية التوسع، وأحدث المستجدات.

يلخص الفصل الختامي المفاهيم الرئيسة الواردة في الكتاب، ويناقش المستجدات الحديثة في هذا المجال والمشكلات الرئيسة التي ما زالت تتطلب إيجاد حل لها، وكذلك بعض التوقعات المستقبلية.

الفصل الثاني
المعالجة اللغوية

٢-١ مقدمة

هناك عدد من المهام اللغوية ذات المستوى المنخفض تشكل أساس خوارزميات معالجة اللغة الأكثر تعقيدًا. في بداية هذا الفصل، سنلقي الضوء على المنهجيات الرئيسة المستخدمة في مهام معالجة اللغات الطبيعية، ومفهوم مسارات مهام معالجة اللغات الطبيعية، وسنقوم بإعطاء أمثلة على بعض الأدوات الرئيسة مفتوحة المصدر. بعد ذلك سنشرح بمزيد من التفصيل المكونات المختلفة للمعالجة اللغوية التي تستخدم عادة في مسارات المهام، كما سنشرح دور هذه المعالجة المسبقة وأهميتها لتطبيقات الويب الدلالي. سنقوم أيضًا بوصف كل عنصر من عناصر مسارات المهام ووظيفته، وسنوضح كيف يرتبط بالمكونات السابقة ويبنى عليها. في كل مرحلة، سنقدم أمثلة على الأدوات وسنقوم بوصف أدائها النموذجي، إلى جانب بعض التحديات والصعوبات المحتملة المرتبطة بكل مكوّن، وسيناقش الفصل الثامن التعديلات المحددة التي يتم إدخالها على هذه الأدوات لتكييفها مع النصوص غير المعيارية مثل نصوص شبكات التواصل الاجتماعي، وتحديدا تويتر.

٢-٢ المنهجيات المتبعة في المعالجة اللغوية

هناك نوعان رئيسان من المنهجيات المتبعة في مهام المعالجة اللغوية: أحدهما منهجية قائمة على المعرفة والآخر منهجية مبنية على التعلم، علما أنه يمكن أيضًا دمجهما معًا. هناك مزايا وعيوب لكل منهجية، ملخصة في الجدول ٢-١.

المنهجية القائمة على المعرفة أو القائمة على القواعد تعدُّ من الأساليب التقليدية بصفة عامة، وقد حلت محلها في كثير من الحالات منهجيات التعلم الآلي نظرًا لأن عملية معالجة كميات هائلة من البيانات بسرعة وكفاءة لم تعد تشكل معضلة بقدر ما كان الأمر عليه في الماضي. تستند المنهجية القائمة على المعرفة على قواعد مكتوبة يدويًّا، وتجري كتابة هذه القواعد عادة على يد متخصصين في مجال معالجة اللغات الطبيعية، وتتطلب معرفة قواعد اللغة والمهارات اللغوية، فضلاً عن امتلاك ملكة البديهة. تكون هذه المنهجيات ذات فائدة أكبر إن أمكن تعريف المهمة بسهولة بواسطة القواعد (على سبيل المثال قاعدة: «الاسم الصحيح -في اللغة الإنجليزيّة- يبدأ دائمًا بحرف

كبير»)، وفي العادة، يمكن استثناء هذه القواعد بسهولة. عندما لا تنطبق القاعدة اللغوية بشكل مباشر تولد هذه المنهجية إشكالية أكبر من السابق (على سبيل المثال: في تغريدات تويتر غالبًا لا يستخدم الناس الأحرف الكبيرة لكتابة الأسماء الصحيحة -في اللغة الإنجليزية-). من بين المزايا الكبيرة للمنهجية القائمة على المعرفة السهولة الكبيرة في فهم النتائج. عندما يتعرف النظام على شيء ما بشكل غير صحيح، يكون بوسع المطور التحقق من القواعد وتحديد سبب حدوث الخطأ، ومن ثم يحتمل أن يكون بمقدوره تصحيح القواعد أو كتابة قواعد إضافية لحل المشكلة. ومع ذلك، يمكن أن تستهلك عملية كتابة القواعد الكثير من الوقت، وفي حال حدوث تغيير في المهمة، فقد يضطر المطوّر إلى إعادة كتابة العديد من القواعد.

منهجيات تعلم الآلة تحظى بشعبية أكبر في الآونة الأخيرة مع ظهور أجهزة قوية ومتطورة، وأيضًا بسبب عدم وجود ضرورة لامتلاك خبرة في المجال المعنِيّ أو امتلاك معرفة لغوية. ولذلك أصبح بالإمكان أن ننشئ نظامًا خاضعًا للإشراف بسرعة كبيرة إذا توفرت بيانات تدريبية كافية، وبوسعنا الحصول على نتائج معقولة بعد تدريب محدود جدًّا. غير أن الحصول على بيانات تدريبية كافية أو إنشاءها غالبًا ما يطرح إشكالية كبيرة للغاية ويستغرق وقتًا طويلا، ولا سيّما إذا كان لا بدّ من القيام بهذه العملية يدويًّا. يعني هذا الاعتماد على بيانات التدريب أيضًا أن التكيّف مع أنواع جديدة من النصوص أو المجالات أو اللغات سيكون مكلفًا على الأرجح، حيث يتطلب توفر قدرٍ كبير من بيانات التدريب الجديدة. لذا، فإن القواعد التي يكون البشر قادرين على قراءتها عادة ما تكون أسهل في التكيّف مع اللغات وأنواع النصوص الجديدة مقارنة بتلك المبنية على أساس النماذج الإحصائية. كما يمكن معالجة مشكلة توفر بيانات التدريب الكافية عبر الدمج بين التعلم الآلي والطرق غير الخاضعة أو شبه الخاضعة للإشراف: هذه الموضوعات ستناقش بشكل موسع في الفصلين الثالث والرابع، مع العلم أنها عادة ما تعطي نتائج أقل دقة مقارنة بنتائج التعلم الخاضع للإشراف.

الجدول ٢-١ : ملخص المنهج القائم على المعرفة في مقابل المنهج القائم على التعلم الآلي في معالجة اللغات الطبيعية

المنهج القائم على المعرفة	أنظمة التعلم الآلي
تقوم على قواعد مكتوبة يدويًّا	تستخدم علم الإحصاء أو أساليب التعلم الآلي الأخرى
جرى تطويرها على يد متخصصين بمعالجة اللغات الطبيعية	لا يتعين على المطورين أن يكونوا على دراية بمعالجة اللغات الطبيعية
تستغل ملكة البديهة البشرية	تتطلب كميات ضخمة من البيانات التدريبية
نتائج سهلة الاستيعاب	يصعب فهم أسباب وقوع الأخطاء
قد تستهلك عملية التطوير وقتًا طويلاً للغاية	عملية التطوير سهلة وسريعة
قد تتطلب التغييرات إعادة كتابة القواعد	قد تتطلب التغييرات عملية إعادة إضافة تعليقات وشروحات

٢-٣ مسارات مهام معالجة اللغات الطبيعية

تتألف مسارات مهام ما قبل معالجة اللغات الطبيعية إجمالا من المكونات التالية، كما هو مبين في الشكل ٢-١ :

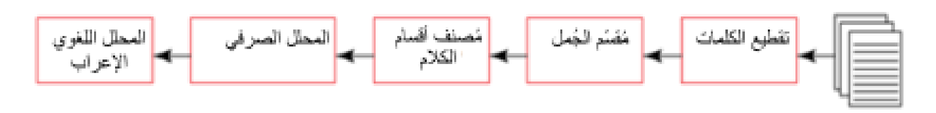
تقطيع الكلمات Tokenization.

تقسيم الجمل Sentence splitting.

تصنيف أقسام الكلام Part-of-speech tagging.

التحليل الصرفي Morphological analysis.

التحليل اللغوي وتجزئة النص Parsing and chunking.



الشكل ٢-١ : نموذج مسارات مهام ما قبل المعالجة اللغوية.

عادة ما تكون المهمة الأولى تجزئة كلمات النص إلى قطع، تليها مهمة تقسيم الجمل، بهدف تقطيع النص إلى وحدات لغوية (تكون في العادة كلمات وأرقام وعلامات ترقيم والمسافات بين الكلمات) وجُمل على التوالي. تضع مهمة تصنيف أقسام الكلام (POS) كل جزء من أجزاء الجملة في فئة نحوية. عند التعامل مع نص متعدد اللغات مثل التغريدات، يمكن إضافة خطوة إضافية تتمثل في التعرف على اللغة قبل أن يتم ذلك، كما سنناقش في الفصل الثامن. التحليل الصرفي ليس إلزاميًّا، لكنه غالبًا ما يُستخدم ضمن مكونات مسارات المهام، ويقوم بشكل أساسي بإيجاد جذر كل كلمة (وهو بذلك يُعدُّ شكلاً أكثر تعقيدًا -إلى حد ما- من مهمة توليد جذع الكلمة أو مهمة التجذير (أي الحصول على جذر الكلمة). أخيرًا، يمكن استخدام أدوات تحليل و/ أو تقطيع أجزاء الكلمة بغية تحليل النص من الناحية التركيبية، وتحديد أمور من قبيل العبارات الاسمية والفعلية في حالة تقطيع النص، أو إجراء تحليل أكثر تفصيلاً للبنية النحوية في حالة التحليل أو الإعراب اللغوي.

فيما يتعلق بمجموعات الأدوات، توفر منصة عمل GATE [4] عددًا من مكونات المعالجة اللغوية المسبقة مفتوحة المصدر بموجب ترخيص LGPL. كما تحتوي على مسارات مهام جاهزة يمكن استخدامها لاستخلاص المعلومات، تسمى ANNIE، وتضم أيضًا عددًا كبيرًا من أدوات المعالجة اللغوية الإضافية مثل مجموعة مختارة من المحللات اللغوية المختلفة. وعلى الرغم من أن منصة GATE توفر خاصية العمل مع المكونات القائمة على التعلم الآلي، إلا أن نظام ANNIE يتبع منهجية مبنية على المعرفة بشكل عام، وهو ما يجعل عملية التكييف سهلة. يمكن إضافة موارد إضافية عن طريق آلية إضافة الملحقات أو المكونات الإضافية، بما في ذلك مكونات من مسارات المهام الأخرى مثل أدوات Stanford CoreNLP. مكونات GATE كلها مبنية بواسطة لغة البرمجة جافا، وهو ما يجعل عملية الدمج سهلة ويجعل المكونات غير محددة بمنصة معينة.

Stanford CoreNLP [5] أداة تضم مسارات مهام مفتوحة المصدر، وهي متاحة بموجب ترخيص GPL، ويمكنها أداء جميع مهام المعالجة اللغوية الأساسية المذكورة في هذا القسم، وذلك عبر واجهة برمجة تطبيقات بسيطة مكتوبة بلغة البرمجة جافا. إحدى

المزايا الرئيسة لهذه الأداة أنه يمكن استخدامها في سطر الأوامر دون الحاجة إلى فهم أطر أكثر تعقيدًا مثل GATE أو UIMA، وهذه البساطة، إلى جانب جودة النتائج العالية عمومًا، هي السبب في جعلها تُستخدم على نطاق واسع عندما تكون المعلومات المطلوبة معلومات لغوية بسيطة مثل علامات تصنيف أقسام الكلام. كما هو الحال مع ANNIE، تعدُّ معظم مكونات Stanford CoreNLP مكونات مبنية على قواعد، باستثناء برنامج تصنيف أقسام الكلام.

OpenNLP[1] أداة مفتوحة المصدر تُستخدم لمعالجة اللغة وتعتمد على التعلم الآلي، وتستخدم الإنتروبيا القصوى maximum entropy والمصنفات المعتمدة على البيرسيبترونز (مستقبلات الشبكات العصبونية الاصطناعية). هذه الأداة متاحة مجانًا بموجب ترخيص Apache. وكما هو الحال مع أداة Stanford CoreNLP، يمكن تشغيل OpenNLP على سطر الأوامر بواسطة واجهة برمجة تطبيقات بسيطة مكتوبة بلغة البرمجة جافا. وعلى الرغم من كون المكونات المختلفة الموجودة في الأجزاء الأخرى ضمن مسارات المهام تعتمد على أجزاء الجمل والجمل بشكل أساسي، مثلما هو الحال مع معظم مسارات المهام الأخرى، لكن يمكن تشغيل مُقسّم النص إما قبل مجزئ الوحدات اللغوية أو بعده، وهو أمر غير معتاد نوعًا ما.

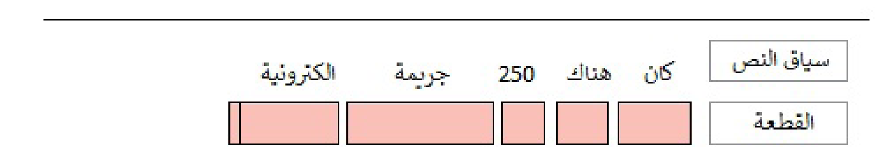
NLTK [6] أداة مفتوحة المصدر مكتوبة بلغة بايثون (python)، وهي متاحة بموجب رخصة Apache، وتحظى بشعبية كبيرة أيضًا بسبب بساطتها وواجهة المستخدم الخطية، خصوصًا عندما لا تكون هناك حاجة لوجود الأدوات المبنية على لغة جافا. توفر هذه الأداة كذلك عددًا من الأشكال المختلفة لبعض المكونات، سواءٌ أكانت مكونات مبنية على القواعد أم مبنية على التعلم الآلي.

في باقي أجزاء هذا الفصل، سنقوم بشرح مكونات مسارات المهام الفردية بمزيد من التفصيل، وذلك باستخدام الأدوات ذات الصلة الموجودة في خطوط الأنابيب كأمثلة.

1- http://opennlp.apache.org/index.html

٢-٤ تقطيع كلمات النص

تجزئة كلمات نص إلى قطع هي مهمة تقسيم النص المُدخل إلى وحدات بسيطة جدًّا، تدعى الوحدات اللغوية (tokens)، وهذه الوحدات تشير عمومًا إلى الكلمات والأرقام والرموز، وعادة ما يتم فصلها بواسطة المسافة البيضاء في اللغة الإنجليزية. تجزئة الوحدات اللغوية خطوة مطلوبة في جميع تطبيقات المعالجة اللغوية تقريبًا، لأن الخوارزميات الأكثر تعقيدًا مثل خوارزميات تصنيف أقسام الكلام، تتطلب في الغالب وجود هذه الوحدات كمدخلات لها، بدلاً من استخدام النص الخام. وبناءً على ذلك، من المهم استخدام مجزئ وحدات لغوية ذي جودة عالية، لأنه من المرجح أن تؤثر الأخطاء على نتائج جميع مكونات معالجة اللغات الطبيعية التي تأتي في مرحلة لاحقة من مراحل مسارات المهام. تشمل أنواع الوحدات اللغوية الشائعة الأرقام والرموز (على سبيل المثال: $ و ٪)، وعلامات الترقيم، والكلمات على اختلاف أنواعها، على سبيل المثال، الكلمات المكتوبة بالأحرف الكبيرة والصغيرة والكلمات المكتوبة بأحرف مختلفة الحالة -في اللغة الإنجليزيّة-. يُظهر الرسم التوضيحي جملة مقطّعة في الشكل ٢-٢، حيث يشير كل مستطيل وردي إلى وحدة لغوية.



الشكل ٢-٢: رسم توضيحي لجملة مجزأة إلى وحدات لغوية.

قد تضيف برامج تجزئة الوحدات اللغوية عددًا من الخصائص التي تصف الوحدة اللغوية. تشمل هذه الخصائص التفاصيل المتعلقة بأسلوب الإملاء (على سبيل المثال: ما إذا كانت حالة الأحرف كبيرة أو لا -في اللغة الإنجليزيّة-)، ومعلومات إضافية حول نوع الوحدة (سواء أكانت كلمة أم رقمًا أم إحدى علامات الترقيم، وما إلى ذلك). كما يمكن للمكونات الأخرى إضافة خصائص إلى تعليقات وشروحات الوحدات اللغوية الموجودة حاليًّا، مثل التصنيف النحوي للوحدة وتفاصيلها الصرفية، وأي تنظيف أو ضبط (مثل تصحيح كلمة خاطئة). سيرد وصف هذه الأمور في الأقسام والفصول اللاحقة. يبين الشكل ٢-٣ وحدة لغوية تشير إلى كلمة جرائم (offences) المذكورة

في المثال السابق مع إضافة بعض الخصائص منها: نوع الوحدة اللغوية هو كلمة، ويبلغ طولها ٨ أحرف –باللغة الإنجليزيّة– وتستخدم الأحرف الصغيرة في طريقة الإملاء.

بشكل عام، تجزئة كلمات نص مكتوب بشكل جيد إلى وحدات لغوية تُعد عملية موثوقة ويمكن إعادة استخدامها، وذلك بسبب كونها ذات طبيعة تميل إلى عدم المحدودية بنطاق أو مجال معين. ومع ذلك، فإن برامج تجزئة الوحدات اللغوية من هذا القبيل ذات الاستخدامات المتعددة تتطلب عادة أن يتم تكييفها لكي تعمل بشكل صحيح مع أشياء مثل الصيغ الكيميائية ورسائل تويتر وغيرها من أنواع النصوص التي تتسم بقدر أكبر من الخصوصية. تشمل الحالات الأخرى غير القياسية الكلمات الموصولة بواصلة في اللغة الإنجليزية، والتي تُعامل من قبل بعض الأدوات كوحدة لغوية واحدة، بينما تعاملها أدوات أخرى على أنها ثلاث وحدات (أي الكلمتان الموصولتان، بالإضافة إلى الواصلة نفسها). تقوم بعض النظم أيضًا بعملية تقطيع للوحدات اللغوية بشكل أكثر تعقيدًا من ذلك، حيث تأخذ بعين الاعتبار تركيبات الأعداد مثل التواريخ والأوقات (على سبيل المثال: التعامل مع ٠٧:٥٦ كوحدة واحدة). هناك أدوات أخرى تترك هذه المهمة لمكونات أخرى في مراحل لاحقة ضمن مسار المعالجة اللغوية، مثل عنصر التعرف على كيانات الأسماء. هناك مسألة أخرى تتعلق بالفاصلة العليا: على سبيل المثال، في الحالات التي يتم فيها استخدام الفاصلة العليا للدلالة على حرف مفقود وتجمع بذلك من الناحية العملية بين كلمتين من دون وجود مسافة بينهما، مثل it’s باللغة الإنجليزية، أو l’homme باللغة الفرنسية. في المقابل، تعاني الأسماء المركبة في اللغة الألمانية من عكس هذه المشكلة، حيث يمكن كتابة العديد من الكلمات معًا من دون مسافة. بالنسبة لمقطعات الوحدات اللغوية الألمانية، فإن وجود وحدة إضافية تقسم التركيبات اللغوية إلى أجزائها المكونة قد يكون مفيدًا جدًّا، ولا سيّما لأغراض الاسترجاع. تعد وحدة التجزئة الإضافية هذه بالغة الأهمية أيضًا لرسم حدود الكلمات عندما يتعلق الأمر بالعديد من لغات شرق آسيا مثل الصينية، التي لا يوجد فيها مفهوم المسافات بين الكلمات.

سياق النص	كان	هناك	250	جريمة	الكترونية
القطعة					

الفئة	NSS - اسم، جمع
النوع	كلمة
الطول	٨ –باللغة الإنجليزيّة
التهجئة	أحرف صغيرة
سلسلة الأحرف	offences – جرائم

الشكل ٢-٣: رسم توضيحي لجملة مجزأة إلى وحدات لغوية.

بسبب كون عملية تقطيع كلمات النص تتبع بشكل عام مجموعة صارمة من القيود التي تحدد ما الذي يشكل وحدة لغوية، إلا أنه كثيرًا ما يجري استخدام أساليب المطابقة القائمة على الأنماط في هذه الأدوات، على الرغم من أن بعض الأدوات تستخدم مناهج أخرى. تعدُّ أداة تجزئة الوحدات اللغوية OpenNLP TokenizerME[1]، على سبيل المثال، مقطع بنظرية التحول نحو الحد الأقصى قابل للتدريب، وتستخدم نموذجًا إحصائيًّا، استنادًا إلى مكنز تدريبي، علمًا أنه يمكن إعادة التدريب باستخدام مكنز جديد.

تعتمد أداة تجزئة الوحدات اللغوية ANNIE Tokenizer[2] الخاصة بمنصة GATE على مجموعة من قواعد التعبيرات القياسية التي يتم ترجمتها بعد ذلك إلى آلة الحالات المحدودة finite-state machine. يختلف هذا المجزِّئ إلى حد ما عن معظم المجزِّئات الأخرى في أن كونه يحقق أقصى حد ممكن من الكفاءة عن طريق إجراء معالجة خفيفة جدًّا، ويوفر مرونة أكبر عن طريق وضع عبء القيام بعمليات المعالجة الأعمق على

1- http://incubator.apache.org/opennlp/documentation/manual/opennlp.html

2- http://gate.ac.uk

المكونات الأخرى في وقت لاحق في مسارات المهام التي تعدُّ أكثر قدرة على التكيف. يستند الإصدار العام لمجزِّئ ANNIE على معيار التشفير الموحد يونيكود[١]، ويمكن استخدامه في أي لغة توجد فيها مفاهيم مماثلة للوحدات اللغوية والمساحات البيضاء الموجودة في الإنجليزية (أي معظم اللغات الغربية). يمكن أيضًا تكييف المقطّع ليلائم لغات مختلفة إما عن طريق تعديل القواعد الموجودة، أو عن طريق إضافة بعض القواعد الإضافية في مرحلة ما بعد المعالجة. بالنسبة للغة الإنجليزية، هناك مجموعة متخصصة من القواعد، وتتعامل هذه القواعد بشكل رئيس مع استخدام الفواصل العليا في كلمات مثل «don't».

تعدُّ PTBTokenizer[٢] أداة تقطيع تتميز بالكفاءة والسرعة وتعطي نتائج قطعية، وتشكل جزءًا من مجموعة أدوات Stanford CoreNLP. وقد صممت هذه الأداة في البداية لمحاكاة أداة التجزئة الخاصة بـ(Treebank 3 (PTB، ومن هنا جاء اسمه. مثل ANNIE، حيث تعمل هذه الأداة بشكل جيد مع اللغة الإنجليزية واللغات الغربية الأخرى، لكنها تعمل بأفضل صورة عند التعامل مع النصوص الرسمية. وعلى الرغم من كونها قطعية النتائج، إلا أنها تستخدم بعض الاستدلالات الجيدة جدًّا، لذلك وكما هو الحال مع المجزئ ANNIE، فإنه يمكن للمقطع PTBTokenizer أن يقرر عندما تكون علامات الاقتباس المفردة جزءًا من الكلمة، وعندما تعني نقطة النهاية أنه تم الوصول إلى حدود الجملة، وما إلى ذلك. كما يمكن أيضًا تخصيصه بشكل كامل، من حيث وجود عددٍ من الخيارات التي يمكن تعديلها.

توجد في أداة[3]NLTK أيضًا العديد من المجزِّئات المماثلة لـANNIE، أحد هذه المجزِّئات يعتمد على التعبيرات القياسية، ونشير إلى أن NLTK مصممة بلغة بايثون.

١- لفهم معيار التشفير الموحد (يونيكود)، انظر: http://www.unicode.org/standard/WhatIsUnicode.html.

2- http://nlp.stanford.edu/software/tokenizer.shtml

3- http://www.nltk.org/

٢-٥ تقسيم الجمل

تمييز الجمل (أو تقسيم الجمل) هي مهمة تقسيم النص إلى الجمل المكونة له، وعادة تشتمل هذه المهمة على تحديد ما إذا كانت علامات الترقيم، مثل نقطة النهاية والفواصل وعلامات التعجب وعلامات الاستفهام، تدل على نهاية الجملة أو على شيء آخر (الكلام المقتبس، الاختصارات، وما إلى ذلك). تستخدم معظم مقطعات الجمل قوائم الاختصارات للمساعدة في تحديد هذا الأمر: تدل نقطة النهاية عادة على نهاية الجملة ما لم تأتِ بعد اختصار مثل السيد. (.Mr)، أو توجد داخل علامات اقتباس. تشمل الأمور الأخرى تحديد بناء الجملة عند استخدام فواصل الأسطر، على سبيل المثال في العنوانات أو في القوائم النُقطية. تختلف مقسّمات الجمل في كيفية تعاملها مع هذه الأمور.

تنشأ حالات أكثر تعقيدًا عندما يحتوي النص على جدَاول أو عنوانات أو معادلات أو غيرها من علامات التنسيق: عادة ما تكون هذه العناصر هي المصدر الأكبر للأخطاء. تتجاهل بعض مقسّمات الجمل هذه الأشياء تمامًا، وتتطلب أن تدل علامات الترقيم على الحدود الفاصلة بين الجمل. كما تستخدم مقسّمات جُمل أخرى سطرين متتاليين جديدين أو الضغط على مفتاح الإدخال (enter /return) كمؤشر على نهاية الجملة، في حين توجد أيضًا حالاتٌ يدل فيها سطرٌ جديد واحد أو ضغطة واحدة على مفتاح الإدخال على نهاية الجملة (على سبيل المثال: التعليقات الموجودة داخل الرموز البرمجية أو القوائم النقطية / المرقمة التي تضم عنصرًا أو مُدخلاً واحدًا في كل سطر). يوفر مقسّم الجمل ANNIE الخاص بمنصة عمل GATE في الواقع عدة بدائل من أجل السماح للمستخدم باتخاذ قرار بشأن الحل الأنسب للنص المحدد الموجود بين يديه. تعدُّ علامات التنسيق في لغة HTML وعلامات التصنيف أو الوسوم (hash tags) المستخدمة في تويتر وبناء الجمل في المواقع التعاونية المعتمدة على مساهمة المستخدمين (wiki)، وغير ذلك من أنواع النصوص الخاصة مشكلة إلى حد ما لمقسّمات الجمل المتعددة الاستخدامات والتي تم تدريبها على مكانز خالية من الأخطاء، كنصوص الصحف. لاحظ أنه في بعض الأحيان يتم إجراء مهمتي تجزيء الجمل وتقسيم الجمل كمهمة واحدة بدلاً من إجرائهما واحدة تلو الأخرى.

تستفيد مقسّمات الجمل في العادة من نصوص سبق تجزيئها. يستخدم مقسّم الجمل ANNIE من GATE المنهج المعتمد على القواعد والمستند بدوره على أنماط كتابة قواعد لغة [7] JAPE GATE's. تستند هذه القواعد كليًّا على المعلومات التي ينتجها مقطّع الوحدات اللغوية وبعض القوائم التي تضم الاختصارات الشائعة، ويمكن تعديلها بسهولة عند الضرورة. تتوفر هذه المقسّمات في صيغ عديدة، كما أوردنا ذلك سابقا.

على عكس ANNIE، يعمل مقسّم الجمل OpenNLP عادة قبل مقطع الوحدات اللغوية، ويستخدم نهج التعلم الآلي، مع كون النماذج المزودة متدربة على نص غير مجزأ إلى وحدات لغوية، على الرغم من أنه من الممكن أيضًا تجزئة النص أولاً، ليقوم مقسّم الجمل بعد ذلك بمعالجة النص المقطع مسبقًا. هناك عيب واحد في مقسّم الجمل OpenNLP وهو عدم قدرته على تحديد الحدود الفاصلة بين الجمل استنادًا إلى محتويات الجملة، ما قد يسبب وقوع أخطاء في المقالات التي لها عنوانات لأنه يتم تحديدها بصورة خاطئة على أنها تشكل جزءًا من الجملة الأولى.

يستخدم NLTK مقسّم الجمل Punkt [8]، حيث يستخدم هذا البرنامج نهجًا مستقل اللغة وغير خاضع للإشراف في تحديد الحدود الفاصلة بين الجمل، استنادًا إلى تحديد الاختصارات والأحرف الأولى والأعداد الترتيبية. خلافًا لمعظم مقسّمات الجمل، لا تعتمد عملية الكشف عن الاختصارات في هذا المقسّم على قوائم تم تجميعها مسبقًا، بل تعتمد بدلاً من ذلك على أساليب الكشف عن المتلازمات اللفظية مثل لوغاريتم الاحتمال (log-likelihood).

تستفيد أداة Stanford CoreNLP من النصوص المجزأة إلى وحدات لغوية ومجموعة من أشجار القرارات الثنائية باتخاذ قرار بشأن مواقع الحدود الفاصلة بين الجمل. وكما هو الحال مع مقسّم الجمل ANNIE، تكمن المشكلة الرئيسة في محاولة اتخاذ قرار فيما إذا كانت نقطة النهاية تدل على نهاية جملة أم لا.

في بعض الدراسات، سجل مقسّم الجمل الخاص بـ Stanford أعلى دقة من بين سائر البرامج الشائعة لتقسيم الجمل، على الرغم من أن الأداء سوف يختلف من حالة لأخرى بالطبع تبعًا لطبيعة النص. تسجل مقسّمات الجمل الحديثة كالتي ذكرت آنفًا أعلى دقة بنسب تتراوح بين ٩٥-٩٨٪ عند العمل على النصوص المكتوبة بشكل جيد.

وكما هو الحال مع معظم أدوات المعالجة اللغوية، يوجد لدى كل مقسّم للجمل نقاط قوة ونقاط ضعف، وهي غالبا ما ترتبط بخصائص محددة في النص؛ على سبيل المثال، قد تعطي بعض مقسّمات الجمل أداءً أفضل عند التعامل مع الاختصارات، في حين قد يكون أداؤها أسوأ عند التعامل مع الكلام المقتبس.

٢-٦ تصنيف أقسام الكلام

يُعنى تصنيف أقسام الكلام (POS) بوضع علامات على الكلمات تشير إلى تصنيف الكلام الذي تنتمي إليه، على سبيل المثال، الأسماء والأفعال والصفات. تنقسم هذه الفئات اللغوية الأساسية عادة إلى أصناف دقيقة، حيث تميز هذه الأصناف على سبيل المثال بين الأسماء المفردة وأسماء الجمع وأزمنة الأفعال. بالنسبة للغات الأخرى غير الإنجليزية، يمكن أيضًا إدراج الجنس في التصنيف. تعدُّ مجموعة التصنيفات الممكنة التي يجرى استخدامها أمرًا بالغ الأهمية وتختلف باختلاف الأدوات المستخدمة في التصنيف، وهو ما يجعل قابلية التشغيل البيني بين الأنظمة المختلفة مهمة صعبة. من بين التصنيفات الشائعة جدًّا في اللغة الإنجليزية Penn Treebank (PTB) [9]؛ وتشمل التصنيفات الشائعة الأخرى تلك المستمدة من مكنز براون (Brown) [10] ومكنز LOB (لانكستر – أوسلو / بيرغن) [11]، على التوالي. يبين الشكل ٢-٤ مثالاً على بعض النصوص المصنفة وفقًا لتصنيف أقسام الكلام، باستخدام تصنيفات مكنز PTB (ملحوظة: لا يحتوي المكنز على اللغة العربية، وهذا المثال بعد ترجمته من اللغة الإنجليزيّة).

سياق النص	كان	هناك	250	جريمة	الكترونية	
القطعة	VBD	EX	CD	NNS	NN	.

الشكل ٢-٤: رسم توضيحي لجملة مصنفة وفقًا لتصنيف أقسام الكلام.

تحديد تصنيف قسم الكلام لا يتم بأخذ الكلمة نفسها في الاعتبار فحسب، بل أيضًا من خلال السياق الذي تظهر فيه، والسبب هو أن العديد من الكلمات غامضة، والرجوع إلى المعجم لا يعدُّ كافيًا لحل هذه المشكلة. على سبيل المثال، يمكن أن تكون

كلمة love [حُب] اسمًا أو فعلاً، بناءً على السياق (جملة «أحب السمك» مقابل جملة «الحب هو كل ما تحتاجه»).

تستخدم أدوات تصنيف أقسام الكلام عادة منهجيات التعلم الآلي، لأنه من الصعب جدًّا وصف جميع القواعد اللازمة لتحديد التصنيف الصحيح في ضوء سياق معين (بالرغم من استخدام الأساليب التي تعتمد على القواعد). تستخدم بعض المنهجيات الأكثر شيوعا ونجاحًا نماذج ماركوف المخفية (HMMs) أو منهجية التحول القصوى. يعدُّ مُصنِف Brill التحويلي الذي يعتمد على القواعد [12]، والذي يستخدم تصنيفات PTB، من المُصنِفات الأكثر شهرة التي تستخدم في العديد من مجموعات أدوات معالجة اللغات الطبيعية الرئيسة. يستخدم مُصنِف Brill معجمًا افتراضيًّا ومجموعة قواعد مستقاة من مجموعة كبيرة من البيانات التدريبية عن طريق التعلم الآلي. وبالمثل، فإن مُصنِف OpenNLP يستخدم أيضًا نموذجًا تم تدريبه من مكنز بهدف التنبؤ بالتصنيف الصحيح لقسم الكلام وفقًا لتصنيفات PTB. يمكن أيضًا تدريبه إما بواسطة التحول القصوى أو بواسطة نموذج معتمد على البيرسيبترونز (Perceptron–based model). يستند مُصنِف Stanford لتحديد أقسام الكلام أيضًا على منهجية التحول الأقصى [13] ويستخدم تصنيفات PTB. يعدُّ مُصنِف TNT (Trigrams'n'Tags) [14] مُصنِفًا إحصائيًّا سريعًا وفعالاً، ويستخدم تطبيق خوارزمية فيتربي (Viterbi) لنماذج ماركوف من الدرجة الثانية.

من ناحية أدوات معالجة اللغات الطبيعية الرئيسة، يوجد لدى بعضها (مثل Stanford CoreNLP) مُصنفات أقسام الكلام الخاصة بها، كما هو موضح أعلاه، في حين يستخدم بعضهم الآخر تطبيقات موجودة بالفعل أو صيغًا مغايرة من هذه التطبيقات. على سبيل المثال، يستخدم NLTK تطبيقات مبنية على لغة بايثون لمُصنِف Brill ومُصنِف ستانفورد ومُصنِف TNT. كما يعدُّ مُصنِف أقسام الكلام الإنجليزي الخاص بنظام ANNIE التابعة لمنصة GATE [15] نسخة معدلة من مُصنِف Brill جرى تدريبه على مكنز كبير مأخوذ من أرشيف صحيفة وول ستريت جورنال. يقوم هذا المُصنِف بإصدار تصنيف لقسم الكلام على شكل إضافة تعليق وشرح لكل كلمة أو رمز. من بين المزايا الكبيرة لهذا المُصنِف إمكانية تعديل المعجم يدويًّا بسهولة عن

طريق إضافة كلمات جديدة أو تغيير قيمة التصنيفات المحتملة المرتبطة بكلمة ما أو ترتيب هذه التصنيفات. يمكن أيضًا إعادة تدريب المُصنِف على مكنز جديد، على الرغم من أن هذا الأمر يتطلب مجموعة كبيرة من النصوص المُصنفة مسبقًا في نطاق/ نوع ذي صلة، وهو ما لا يمكن إيجاده بسهولة.

عادة ما تكون دقة هذه المُصنِفات متعددة الاستعمالات والتي يمكن إعادة استخدامها ممتازة (98-97٪) عندما تُستخدم مع نصوص مماثلة لتلك التي تم تدريب المُصنِفات عليها (المقالات الإخبارية في الغالب). ومع ذلك، فإن الدقة يمكن أن تضعف بشكل كبير جدا عند تعاملها مع مجالات وأنواع جديدة من النصوص، أو بيانات تحوي قدرا أكبر من التشويش، مثل نصوص شبكات التواصل الاجتماعي، وهو ما قد يترك تأثيرًا خطيرًا على العمليات الأخرى التي تأتي لاحقًا ضمن مسارات المهام، مثل تمييز كيانات الأسماء، وتعلم الكيانات المعجمية عن طريق الأنماط المعجمية النحوية، واستخلاص العلاقات والأحداث، وحتى مهام تعدين الآراء، وكلها تحتاج إلى تصنيفات لأقسام الكلام يمكن الوثوق بها لكي تعطي نتائج عالية الجودة.

٢-٧ التحليل الصرفي

يتعلق التحليل الصرفي بشكل أساسي بالتعرف على الوحدات اللغوية داخل الكلمة وتصنيفها، ويتم عادة تجزئة الكلمة إلى الجذر مع السوابق واللواحق، على سبيل المثال، يتكون الفعل walked من الجذر walk واللاحقة ed-. ينطبق التحليل الصرفي في اللغة الإنجليزية على الأفعال والأسماء، والسبب هو أن الأفعال والأسماء قد تظهر في النص في صيغة أشكال مختلفة تنشأ بفعل الصرف الإعرابي. يشير مصطلح الصرف الإعرابي إلى الأشكال المختلفة للكلمات التي تعكس المزاج وأزمنة الفعل والعدد وما شابه، مثل صيغة الماضي لفعل ما أو صيغة الجمع لاسم معين. يظهر الصرف في اللغة الإنجليزية عادة عن طريق إضافة لاحقة إلى جذر الكلمة (على سبيل المثال: walk، walked، box، boxes) أو عن طريق التعديلات الداخلية الأخرى مثل تغيير الحروف المتحركة (على سبيل المثال: run، ran، goose، geese). في اللغات الأخرى، يمكن استخدام السوابق (إضافة مقطع في بداية الكلمة) أو المتوسطات (إضافة مقطع في وسط الكلمة)، إلى

جانب تغييرات أخرى. تعرض بعض أدوات التحليل الصرفي هذه التعديلات الداخلية على شكل تمثيلات بديلة للَّاحقة الافتراضية. نعني بذلك أنه إذا كانت صيغة الجمع لاسم ما تُعرض عادة بإضافة اللاحقة –s فإن الصيغة التي تعرضها أداة التحليل الصرفي ستكون اللاحقة –s حتى في حال صيغ الجمع من قبيل geese. من الناحية الفعلية، تعامل الأداة ببساطة الصيغة التي طرأ فيها تغيير غير اعتيادي على الحروف المتحركة كنوع من المتغير السطحي التمثيلي للسابقة أو اللاحقة المعيارية [أي اللاحقة المستخدمة عادة وهي إضافة s– في نهاية الكلمة]. على سبيل المثال، يعرض المحلل الصرفي الخاص بمنصة GATE كلمة geese على أنها مكونة من الجذر goose واللاحقة s–.

في العادة، تتعامل أدوات معالجة اللغة الطبيعية التي تقوم بإجراء التحليل الصرفي مع الصرف الإعرابي فقط، كما شرحنا أعلاه، لكنها لا تقوم بإجراء الصرف الاشتقاقي. الاشتقاق هو عملية استخراج أصغر وحدات لغوية ذات معنى (morphemes)، وهو ما ينشئ كلمة جديدة من الكلمات الموجودة، وعادة يشمل ذلك تغييرًا في التصنيف النحوي (على سبيل المثال: إنشاء الاسم worker [عامل] من الفعل work [عمل]، أو الاسم loudness [صخب] من الصفة loud [صاخب]).

في كثير من الأحيان، تكون أدوات التحليل الصرفي في اللغة الإنجليزية معتمدة على القواعد، وذلك لأن غالبية الأشكال الإعرابية تتبع قواعد وأنماطًا نحوية (على سبيل المثال: أسماء الجمع تُنشأ عادة عن طريق إضافة –s أو –es في نهاية صيغة المفرد). يمكن أيضًا معالجة الاستثناءات بسهولة كبيرة بواسطة القواعد، كما يمكن الافتراض أن الكلمات المجهولة تتبع القواعد الافتراضية. المحلل الصرفي في منصة عمل GATE مبني على القواعد، حيث تدعم لغة القواعد (flex) القواعد والمتغيرات التي يمكن استخدامها في التعابير النمطية. يمكن أيضًا أخذ بطاقات تصنيف أقسام الكلام في الحسبان إن كان ذلك مرغوبًا فيه، وهذا يعتمد على عامل الإعداد. تكون مُدخلات المحلل الصرفي على شكل مستند مجزأ، ويقوم بتحليل وحدة لغوية واحدة إلى جانب بطاقة تصنيف أقسام الكلام الخاصة بها في كل مرة، ومن ثمّ يحدد جذر الكلمة وكذلك السابقة أو اللاحقة المضافة إليها. بعد ذلك تُضاف هذه القيم إلى بطاقة تصنيف أقسام الكلام كخصائص.

تستخدم أداة Standford الصرفية أيضًا منهجية معتمدة على القواعد، وتستند على محوّل آلات محدودة (finite-state transducer)، وهي مكتوبة بلغة flex. لكنها وبعكس أداة GATE الصرفية، تتطلب استخدام بطاقات تصنيف أجزاء الكلام بالإضافة إلى الوحدات اللغوية، كما أنها يتولد منها كلمات من دون زوائد وترجع إلى أصلها المعجمي (lemmas) بدلاً من أن تكون على شكل سوابق ولواحق.

توفر NLTK تطبيقًا لتحليل لغوي يعتمد على خاصية morphy المدمجة في نظام WordNet. WordNet [16] هو عبارة عن قاعدة بيانات معجمية إنجليزية شبيهة بقاموس أو موسوعة مفردات، حيث يتم تصنيف الأسماء والأفعال والصفات وظروف الأحوال إلى مجموعات من المترادفات المعرفية (Synsets)، تعبر كل واحدة منها عن فكرة أو مفهوم معين. ترتبط المترادفات المعرفية بواسطة علاقات معرفية- دلالية ومعجمية. صُممت خاصية morphy لكي تتيح للمستخدمين البحث عن شكل صرفي لكلمة ما مقارنة بشكلها الجذري المدرج في قاعدة بيانات WordNet المعجمية، وتتبع أسلوبًا مبنيًّا على القواعد يضم قوائم تحتوي على نهايات صرفية أو إعرابية، وذلك استنادًا إلى التصنيف النحوي للكلمة، كما تستخدم قائمة استثناءات خاصة بكل تصنيف نحوي يتم البحث فيها عن الصيغة الصرفية. وكما هو الحال مع أداة Stanford، تكون نتيجة البحث عبارة عن جذر الكلمة فقط وليس السابقة أو اللاحقة. أضف إلى ذلك أنها قادرة على معالجة الكلمات الموجودة داخل معجم WordNet فقط.

لا توفر OpenNLP في الوقت الراهن أي أدوات لإجراء التحليل الصرفي.

٢-٧-١ اشتقاق جذع الكلمة

تنتج أدوات اشتقاق جذع الكلمة الشكل الجذعي لكل كلمة، على سبيل المثال تشترك الكلمتان driving وdrivers في الجذع drive، فيما يميل التحليل الصرفي إلى إنتاج الأشكال الجذرية للكلمات إضافة إلى سوابقها و/ أو لواحقها، على سبيل المثال drive و driver للأمثلة السابقة، إضافة إلى اللاحقتين ing- وs- على التوالي. هناك حيرة كبيرة حول الفرق بين توليد جذع الكلمة والتحليل الصرفي، وذلك بسبب التباينات الكبيرة التي يمكن أن توجد بين أدوات توليد جذع الكلمة في طريقة عملها وفي البيانات الصادرة منها. بصفة عامة، لا تحاول أدوات توليد جذع الكلمة إجراء تحليل لأصل

أو جذع الكلمة ولاحقتها، بل تقوم ببساطة بتجريد الكلمة من لاحقتها وإرجاعها إلى الجذع. تتمثل الطريقة الرئيسة التي تختلف فيها أدوات توليد جذع الكلمة بعضها عن بعض في وجود أو غياب الشرط المقيّد الذي يتطلب أن يكون الشكل الجذعي عبارة عن كلمة حقيقية موجودة في اللغة المعنية. تقوم عملية توليد جذع الكلمة الأساسية بإزالة اللاحقة، على سبيل المثال، تتم إزالة اللاحقة ing- من كلمة driving لتصبح -driv. في أغلب الأحيان، لا يتم الإبقاء على التمييز بين الأفعال والأسماء، لذا تُزال اللاحقتان من كلمتي driver وdriving لتتحول كلتاهما إلى الشكل الجذري -driv. تستغل أنظمة استرجاع المعلومات (IR) في الغالب هذا النوع من إزالة اللواحق، وذلك لأنه يمكن إتمامه بواسطة خوارزمية بسيطة ولا يتطلب مهام المعالجة اللغوية الأخرى كتصنيف أجزاء الكلام. تعدُّ عملية اشتقاق جذع الكلمة مفيدة لأنظمة استرجاع المعلومات نظرًا لكونها تجمع بين الأشكال المعجمية-النحوية لكلمة ما تشترك جميعًا في المعنى (وبذلك يصبح بالإمكان استخدام صيغة المفرد أو صيغة الجمع خلال عملية البحث، لتتطابق نتيجة البحث مع إحدى الصيغتين داخل صفحة الويب). لاحظ أنه وخلافًا لمعظم أدوات التحليل الصرفي، يمكن أن تأخذ أدوات اشتقاق جذع الكلمة في الحسبان الأشكال الناشئة عن عمليات الصرف الاشتقاقي، وذلك لأنها تتجاهل الفئة النحوية للكلمة. هناك فرق آخر، وهو أن أدوات توليد جذع الكلمة لا تنظر إلى السياق المحيط بالكلمة، بل تنظر فقط إلى الكلمة وحدها بمعزل عن السياق، بينما يمكن أن تأخذ أدوات التحليل الصرفي السياق بعين الاعتبار أيضًا.

يبيّن الشكل ٢-٥ مثالاً يدل على الطرق المحتملة التي يمكن أن تختلف فيها عملية توليد جذع الكلمة عن التحليل الصرفي. تقوم أداة توليد جذع الكلمة الموجودة في منصة عمل GATE بإزالة اللاحقة الاشتقاقية ness- وهو ما يختزل صيغة الاسم loudness في صيغة الصفة loud، كما يتضح من خاصية stem (الجذع) في الجدول أدناه. على الجانب الآخر، لا تهتم أداة التحليل الصرفي بالصرف الاشتقاقي، وتدع الكلمة كما هي بالكامل، كما هو موضح في خاصية root) الجذرloudness) من دون إنتاج أي لاحقة.

The loudness of the music was intolerable

اللاحقة	
الفئة	NNS
النوع	كلمة
الطول	٨
التهجئة	أحرف صغيرة
الجذر	loudness
الجذع	loud
سلسلة الأحرف	loudness

الشكل ٢-٥: مقارنة بين توليد جذع الكلمة والتحليل الصرفي في منصة عمل GATE.

قد تختلف خوارزميات إزالة اللواحق في نتائجها لأسباب عدة. أحد هذه الأسباب يتمثل فيما إذا كانت الخوارزمية تتطلب أن تكون الكلمة الناتجة كلمة حقيقية موجودة في اللغة المعنية. لا تتطلب بعض المنهجيات أن تكون الكلمة موجودة في واقع الأمر في معجم اللغة (ونقصد به جميع الكلمات الموجودة في اللغة).

تعدُّ منهجية Porter Stemmer [17] أشهر خوارزميات توليد جذع الكلمة، وقد صممت بصيغ وأشكال عديدة. ونظرًا للمشكلات التي نجمت عن إنشاء أشكال عديدة لهذه الخوارزمية، فقد ابتكرت Porter لاحقًا لغة Snowball، وهي لغة معالجة صغيرة مصممة خصيصًا لغرض إنشاء خوارزميات توليد جذع الكلمات المستخدمة في عملية استرجاع المعلومات. ومنذ ذلك الوقت، تم استخدام لغة Snowball لإنشاء أدوات متنوعة ومفيدة ومفتوحة المصدر لتوليد جذع الكلمات للعديد من اللغات. توفر منظومة GATE مظلة لعدد من هذه الأدوات، وتضم هذه المظلة 11 لغة من اللغات

الأوروبية، بينما توفر NLTK تطبيقًا لهذه الأدوات للغة بايثون. ونظرًا لكون أدوات توليد جذع الكلمات مبنية على منهجية تعتمد على قواعد ولسهولة تعديلها وفقًا لمنهجية Porter الأصلية، فهذا مما يسهل دمج هذه الأدوات مع المكونات الأخرى ذات المستوى المنخفض التي سبق شرحها في هذا الفصل. تجدر الإشارة إلى أن منظومتي OpenNLP وStanford CoreNLP لا توفران أي أدوات لتوليد جذع الكلمة.

٢-٨ التحليل النحوي

يُعنى التحليل النحوي بتحليل الجمل، وذلك باشتقاق بنيتها النحوية وفقًا للقواعد النحوية. عملية التحليل تشرح بشكل أساسي كيف ترتبط العناصر المختلفة في الجملة بعضها ببعض، على سبيل المثال كيف يتصل الفاعل والمفعول به في فعل معين بعضهما ببعض. هناك الكثير من النظريات النحوية المختلفة في علم اللغويات الحاسوبية، حيث تطرح هذه النظريات أنواعًا مختلفة من البنى النحوية. لهذا السبب، قد تختلف أدوات التحليل بعضها عن بعض، ليس من حيث الأداء فحسب، بل أيضًا من حيث نوع التمثيل الشكلي الذي تُنتجه، وذلك بناءً على النظرية النحوية التي تستخدمها.

تتوفر عدة أدوات تحليل مجانا وتغطي نطاقًا واسعًا وتشمل محلل التبعية Minipar[1]، وكذلك محلل RASP الإحصائي [18] ومحلل Stanford الإحصائي [19]، ومحلل SUPPLE متعدد الاستعمالات [20]. تتوفر جميع هذه الأدوات داخل منصة عمل GATE، وهو ما يعني أن بوسع المستخدم تجربتها جميعًا ومن ثمّ تحديد الأداة الأكثر مناسبة لاحتياجاته.

يُعدُّ محلل Minipar محلل تبعية، بمعنى أنه يحدد علاقات التبعية القائمة بين الكلمات الموجودة في جملة معينة. يقوم هذا المحلل بمعالجة النص جملة بجُملة، ولذا فإنه لا يحتاج سوى إلى مقطّع الجمل كشرط أساسي. يعمل هذا المحلل على أساس تحديد البنى اللغوية وأجزاء الكلام، مثل apposition وجُمل الوصل والفاعل والمفعول به في فعل معين، وكذلك المُحددات، وطريقة ارتباط بعضها ببعض. البدل هي التركيبة اللغوية التي تشير فيها عبارتان اسميتان يوجد بعضهما بجانب بعض إلى الشيء نفسه، على

1- http://www.cs.ualberta.ca/~lindek/minipar.htm

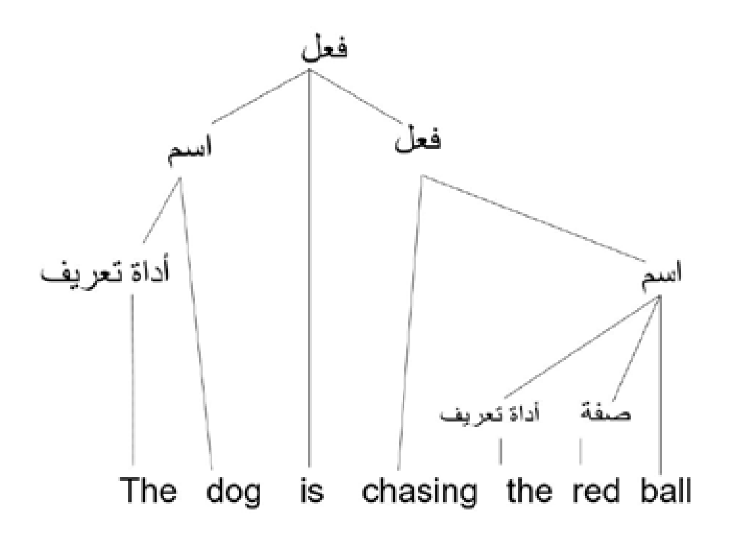
سبيل المثال"my brother John" (أخي جون) أو"Paris, the capital of France" (باريس، عاصمة فرنسا). أما جُمل الوصل فهي تبدأ عادة بأحد ضمائر الوصل (مثل "who" و"which" ...الخ)، وتُدخل تعديلاً على اسم سابق، على سبيل المثال «who was wearing the hat» (الذي كان يرتدي القبّعة) في الجملة «the man who was wearing the hat» (الرجل الذي كان يرتدي القبّعة).

على عكس علاقات التبعية، تعدُّ محللات المكونات مبنية على مفهوم علاقات المكونات، وقد تتضمن عددًا من نظريات القواعد النحوية المختلفة الخاصة بالمكونات، مثل القواعد النحوية الخاصة ببنية العبارات والقواعد النحوية المصنفة والقواعد النحوية المعجمية الوظيفية، وغيرها. تعدُّ علاقة المكونات علاقة هرمية، وهي مستقاة من تقسيم الجملة إلى فاعل ومفعول به في قواعد النحو في اللغتين اللاتينية واليونانية، حيث يتم تقسيم البنية الأساسية للجُملة إلى قسمين هما الفاعل (شبه الجملة الاسمية) والمفعول به (شبه الجملة الفعلية). بعد ذلك تجري تقسيمات إضافية لهذين القسمين كليهما في مستويات تفصيلية أخرى.

يُعدُّ محلل المكوناتShift-Reduce Constituency Parser مثالاً جيدًا على محللات المكونات، ويشكل هذا المحلل جزءًا من أدوات Standford CoreNLP[1].
ظلت عمليات محلل Shift-and-reduce تُستخدم لوقت طويل في عمليات تحليل التبعية بسرعة عالية ودقة فائقة، لكن لم تُستخدم هذه العمليات إلا في الآونة الأخيرة في تحليل المكونات. يهدف محلل Shift-Reduce إلى تحسين عمل محللات المكونات القديمة التي كانت تستخدم خوارزميات تعتمد على الرسوم البيانية (البرمجة الديناميكية) من أجل العثور على نتيجة البحث التي تحصل على أعلى درجة، وكانت هذه المحللات دقيقة وبطيئة للغاية في الوقت نفسه.

يبين الشكل ٢-٦ شجرة تحليل جرى إنتاجها باستخدام القواعد النحوية التبعية، بينما يبين الشكل ٢-٧ شجرة ناتجة عن استخدام القواعد النحوية الخاصة بالمكونات للجملة نفسها (يطارد الكلبُ الكرة الحمراء).

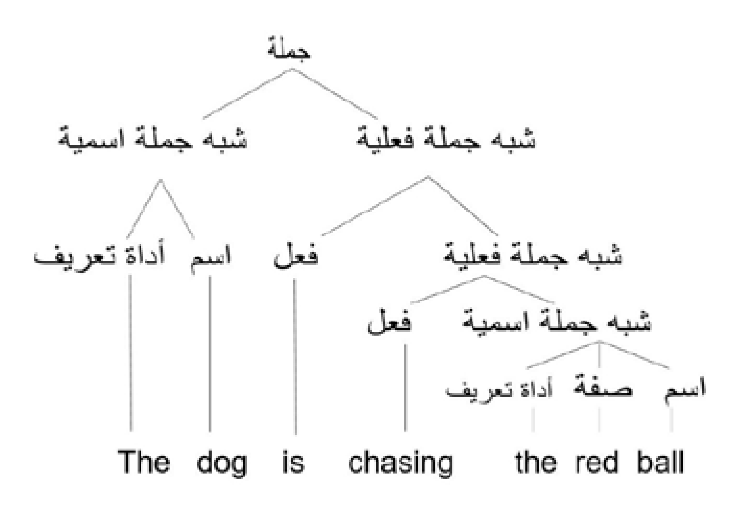
1- http://nlp.stanford.edu/software/srparser.shtm



الشكل ٢-٦: شجرة تحليل تبين علاقات تبعية.

يعدُّ محلل RASP الإحصائي [18] محللاً ذا نطاق حر يتميز بالفاعلية، وهو مصمم للعمل باللغة الإنجليزية. هذا المحلل مرفق بمجزئ وحدات لغوية خاص به، إلى جانب مصنف لأجزاء الكلام ومحلل صرفي خاصين به، وكما هو الحال مع محلل Minipar، يتطلب هذا المحلل أن يكون النص مقطعًا مسبقًا إلى جُمل. محلل RASP متاح بموجب ترخيص LGPL ولذا يمكن استخدامه في التطبيقات التجارية.

يعد محلل Stanford الإحصائي [19] عبارة عن نظام تحليل نحوي قائم على الاحتمالات. يوفر هذا المحلل إما مُخرجات تبعية أو مُخرجات تكون على شكل بنية عبارات أو شبه جُمل. يمكن معاينة النوع الأخير من المُخرجات داخل واجهة المستخدم الرسومية الخاصة بالمحلل، أو عبر استخدام واجهة المستخدم الخاصة بمنصة عمل GATE Developer. يأتي محلل Stanford مرفقًا بملفات بيانات لتحليل لغات تشمل العربية والصينية والإنجليزية والألمانية، وهو مرخص بموجب ترخيص GNU GPL.



الشكل ٢-٧: شجرة تحليل تبين علاقات المكونات.

يُعدُّ محلل SUPPLE محللاً نحويًّا يعمل وفقًا لمفهوم من الأسفل إلى الأعلى bottom-up وهو قادر على إنتاج تمثيل دلالي للجُمل يُسمى النموذج شبه المنطقي المبسط (SQLF). يتميز هذا المحلل بميزة الفاعلية الفائقة، وذلك بفضل قدرته على إصدار نتائج نحوية ودلالية جزئية، وهو ما يجعله قابلاً للتطبيق بصفة خاصة في اشتقاق الخصائص الدلالية لعملية استخلاص العلاقات الدلالية، بناءً على أسلوب التعلم الآلي، لكميات كبيرة من النصوص الحقيقية.

٢-٩ تجزئة النص

تكون خوارزميات التحليل باهظة التكاليف من الناحية الحسابية في كثير من الأحيان، وكما هو الحال مع العديد من أدوات التحليل، تميل هذه الخوارزميات للعمل في أحسن صورها عندما يكون النص الذي تعالجه مشابهًا للنص الذي سبق تدريبها عليه. وبسبب كون مهمة تجزئة النص أكثر تعقيدًا من بعض مهام المعالجة ذات المستوى المنخفض، مثل مهمتي تجزئة وتقسيم الجمل، يكون أداؤها أدنى بكثير في العادة، وهو ما يمكن أن تكون له تداعيات على أي مهمة أخرى من مهام المعالجة التي تأتي لاحقًا، مثل مهمة التعرف على كيانات الأسماء ومهمة إيجاد العلاقات. لهذا السبب، يكون من الأفضل أحيانا التضحية بالمعرفة الإضافية التي يوفرها المحلل مقابل الحصول

على أداة أخف يمكن الاعتماد عليها، مثل أداة تجزئة النص التي تقوم بإجراء تحليلٍ لغوي سطحي -غير عميق-. تتعرف أدوات التقطيع، التي تُعرف أحيانًا بالمحللات السطحية، على سلاسل متتابعة من الكلمات المترابطة مثل أشباه الجمل الاسمية، لكنها وخلافًا للمحللات لا تقدم تفاصيل عن بنيتها الداخلية أو دورها في الجملة.

يمكن تقسيم أدوات تجزئة النص إلى مجزئات العبارات الاسمية ومجزئات العبارات الفعلية. تقل الاختلافات بين هذين النوعين من أدوات التجزئة عن الاختلافات بين خوارزميات التحليل، وذلك لأن عملية التحليل تتم على مستوى تحليل المكونات الرئيسة بشكل إجمالي (coarse-grained level) حيث تقوم أدوات تقطيع الجمل بالتعرف على «أجزاء» النص ذات الصلة، لكنها لا تسعى إلى تحليل تلك الأجزاء. غير أنها قد تختلف فيما بينها فيما تعتبره ذا صلة بجزء النص قيد التحليل. على سبيل المثال، قد تتكون عبارة اسمية بسيطة من سلسلة متتالية تحتوي على مُحدّد اختياري، وصفة أو نعت اختياري واحد أو أكثر، إلى جانب اسم واحد أو أكثر، كما هو مبين في الشكل ٢-٨. من جهة أخرى، قد تتضمن العبارات الاسمية الأكثر تعقيدًا -بالإضافة إلى ما سبق- شبه جملة جار ومجرور أو جُملة وصل تقوم بإدخال تعديل على العبارة الاسمية. تتضمن بعض مجزئات النص هذه الأشياء كجزء من العبارة الاسمية، بينما لا يتضمنها بعضها الآخر (الشكل ٢-١٠). تعتمد عملية اتخاذ قرار بشأن تضمين شبه جملة جار ومجرور أو جُملة وصل في الجملة الاسمية اعتمادًا كبيرًا على الغرض الذي سيتم استخدام أجزاء النص من أجله لاحقًا. على سبيل المثال، إذا كانت أجزاء النص ستُستخدم كمُدخلات لأداة تتعرف على المصطلحات، فينبغي الأخذ بعين الاعتبار ما إذا كان احتمال وجود عبارة تحتوي على شبه جملة جار ومجرور أمرًا ذا صلة أم لا. عندما يتعلق الأمر بتوليد الانطولوجيات، ليست مثل هذه العبارة مطلوبة على الأرجح، لكنها قد تكون مفيدة عند استخدامها كهدف لعملية تحليل المشاعر.

Context	The old man bought a hat.
NounChunk	

الشكل ٢-٨: تقطيع بسيط لشبه جُملة اسمية -الرجل المسن اشترى قبعة.

Context The old man bought a hat with a brim.

NounChunk

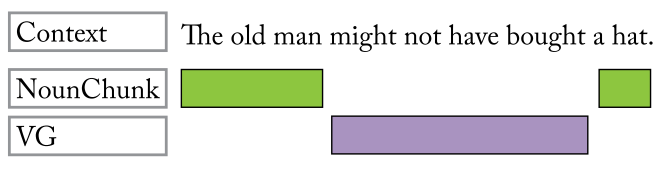
الشكل ٢-٩: تقطيع مركب لشبه جُملة اسمية لا يشمل أشباه جمل الجار والمجرور- الرجل المسن اشترى قبعة ذات حد.

Context The old man bought a hat with a brim.

NounChunk

الشكل ٢-١٠: تقطيع مركب لشبه جُملة اسمية يشمل أشباه جمل الجار والمجرور.

تقوم مجزئات أشباه الجمل الفعلية برسم حدود الأفعال، حيث يمكن أن تتكون الأفعال من كلمة واحدة مثل bought (اشترى) أو مجموعة أكثر تعقيدًا تضم أفعالَ صيغةِ المصدر والأفعال الشكلية المُساعدة وما شابه (على سبيل المثال might have bought [يحتمل أنه اشترى] أو to buy [ليشتري]). قد تتضمن أيضًا عناصر نفي مثل might not have bought (يحتمل أنه لم يشترِ) أو didn't buy (لم يشترِ). يبين الشكل ٢-١١ مثالاً على أحد مُخرجات برنامج لتجزيء الجمل يجمع بين عمليتي تجزيء أشباه الجمل الاسمية وتجزيء أشباه الجمل الفعلية.



الشكل ٢-١١: تقطيع مركب للعبارات الفعلية.

توفر بعض الأدوات أيضًا مهام إضافية، على سبيل المثال يتميز مُصنف أجزاء الكلام TreeTagger [21] (المدرب على قاعدة بيانات Penn Treebank) بقدرته على توليد أجزاء أشباه جُمل الجار والمجرور وأشباه جُمل الصفات وأشباه جُمل ظروف الأحوال وما شابه. قد تكون هذه المهام مفيدة لبناء تمثيل شكلي للعبارة بأكملها من دون الحاجة إلى إجراء تحليل كامل.

وكما رأينا سابقًا، فإن أدوات المعالجة اللغوية ليست خالية من الأخطاء، حتى لو افترضنا أن المكونات التي تعتمد عليها قد قامت بتوليد مُخرجات مثالية. قد يبدو من السهل إنشاء مجزئ لأشباه الجمل الاسمية يعتمد على قواعد نحوية تشمل بطاقات تصنيف لأجزاء الكلام، لكن هذه العملية معرضة للوقوع في الأخطاء بسهولة. دعنا ننظر إلى الجملتين I gave the man food (أعطيتُ الرجل طعامًا) وI bought the baby food (اشتريتُ طعام الطفل). في حالة الجملة الأولى، الرجل والطعام هما عبارتان اسميتان، وهما المفعول به المباشر والمفعول به غير المباشر على التوالي في الفعل gave (أعطيتُ). بإمكاننا إعادة صياغة هذه الجملة لتصبحI gave food to the man (أعطيت الطعام للرجل) من دون حدوث أي تغيير في المعنى، حيث يتضح أن أشباه الجمل الاسمية هذه مستقلٌّ بعضها عن بعض. لكن في المثال الثاني قد تكون شبه الجملة the baby food (طعام الطفل) إما شبه جُملة اسمية فردية تحتوي على الاسم المركب baby food (طعام الطفل) أو تتبع نفس بنية المثال السابق I bought food for the baby (اشتريت طعامًا للطفل). لن يستطيع مجزئ أشباه جُمل اسمية يستخدم نمط «محدد + اسم + اسم» الذي يبدو منطقيًّا التمييز بين هاتين الحالتين. وفي هذه الحالة، قد يكون أداء نموذج معتمد على التعلم أفضل من أداء منهج معتمد على القواعد.

توفر منصة عمل GATE تطبيقات لمقطّعات عبارات اسمية وعبارات فعلية. يعدُّ مجزئ العبارات الاسمية تطبيقًا يعتمد على لغة جافا لمجزئ Ramshaw and Marcus BaseNP [22]، وهو مجزئ مبني على بطاقات تصنيف أجزاء الكلام الخاصة بها، ويستخدم منهج التعلم المعتمد على التحوّل. تكون مُخرجات هذه النسخة من مجزئ العبارات الاسمية مطابقة لمُخرجات النسخة الأصلية المبنية بواسطة لغة Perl/ C++.

مجزئ GATE VP مكتوب بلغة JAPE، وهي لغة خاصة بمنصة عمل GATE تعتمد على كتابة القواعد. هذا المجزئ مبني على أساس قواعد النحو في اللغة الإنجليزية [23، 24]. يتضمن هذا المجزئ قواعد للتعرف على مجموعات الأفعال غير المتكررة، حيث يضم الأفعال المحدودة (is investigating [يُحقق في]) وغير المحدودة (to investigate [التحقيق في]) والنعوت الفعلية (investigated [جرى التحقيق في]) والتراكيب الفعلية الخاصة (is going to investigate [سوف يُحقق في]). جميع

هذه الأشكال الكلمات وأشباه الجمل الظرفية والعبارات السلبية ممكن أن تشمل بهذا المجزئ. ومن مزايا هذه الأداة تحديدها بوضوح لصيغة النفي في الأفعال (مثال don't)، وهو أمر مفيد جدًّا للمهام الأخرى مثل مهمة تحليل المشاعر. تعتمد القواعد على بطاقات تصنيف أجزاء الكلام، إلى جانب بعض الترادفات المحددة (مثال: يمكن استخدام كلمة might للتعرف على الأفعال الشكلية المُساعدة).

يستخدم المجزئ الخاص بمنصة عمل OpenNLP نموذجًا باللغة الإنجليزية مسبق التجهيز ويقوم على منهجية التحول القصوى. وعلى عكس منصة عمل GATE التي يعدُّ المجزِّئان الخاصان بها مستقلين، فإن هذا المحلل يقوم بتحليل النص جملة بجملة، ويقوم بإنتاج أجزاء للعبارات الاسمية والعبارات الفعلية على حد سواء دفعة واحدة، وذلك اعتمادًا على بطاقات تصنيف أجزاء الكلام الخاصة بأشباه الجمل. يتميز مقطّع OpenNLP بسهولة عملية إعادة تدريبه، وهو ما يُسهل بدوره عملية تكييفه مع المجالات وأنواع النصوص الجديدة إذا توفر مكنز ملائم سبق إضافة التعليقات والشروحات إليه.

لا توفر منصتا NLTK Stanford وCoreNLP أي مجزئات للنصوص، على الرغم من إمكانية إنشاء تلك المقطّعات باستخدام القواعد و/ أو تقنية التعلم الآلي من المكونات الأخرى (مثل بطاقات تصنيف أجزاء الكلام) في مجموعة الأدوات ذات الصلة.

٢-١٠ خلاصة

في هذا الفصل، عرضنا مفهوم خط أنابيب معالجة اللغة الطبيعية، وقدمنا شرحًا لمكوناته الرئيسة، مع الإشارة إلى بعض الأدوات ذات المصدر المفتوح المستخدمة على نطاق واسع. من المهم الإشارة إلى أنه في حين يعدُّ أداء مهام المعالجة اللغوية ذات المستوى المنخفض مرتفعًا بشكل عام، إلا أن الأدوات تختلف في أدائها، ولا ينحصر ذلك في دقتها فحسب، بل يشمل أيضًا الطريقة التي تؤدي فيها المهام وفي مُخرجاتها كذلك، وذلك بسبب اتباعها نظريات لغوية مختلفة. لذا من المهم عند اختيار أدوات المعالجة المسبقة فهم ما هي متطلبات الأدوات الأخرى الموجودة في المراحل الفرعية التي تأتي لاحقًا ضمن التطبيق. وعلى الرغم من إمكانية الجمع بين بعض الأدوات (لا سيّما في منصات

عمل من قبيل منصة GATE والمنصات المشابهة لها التي صُممت بالذات لكي تكون قابلة للتشغيل المتبادل)، إلا أن مسألة التوافق بين المكونات المختلفة قد تسبب بعض المشكلات. يعدُّ هذا الأمر من الأسباب التي أدت إلى وجود مجموعات أدوات مختلفة توفر مجموعات أدوات متشابهة لكنها يختلف بعضها عن بعض بشكل طفيف. من المهم كذلك إدراك أثر تغيير المجال ونوع النص من ناحية الأداء، وما إذا كانت الأدوات سهلة التعديل أم لا إن كان الأمر يتطلب ذلك. قد تنشأ مشكلة ما -على وجه الخصوص- بسبب الانتقال من أدوات مُدرّبة على النصوص الإخبارية العادية إلى معالجة نصوص شبكات التواصل الاجتماعي، وهو ما سنناقشه بالتفصيل في الفصل الثامن. وبالمثل، يمكن تكييف بعض الأدوات لتتلاءم مع اللغات الجديدة (وبالأخص المكونات الأولى في سلسلة المعالجة من قبيل مجزئات الوحدات اللغوية)، في حين قد يكون من الصعب تكييف الأدوات الأكثر تعقيدًا من قبيل المحللات اللغوية مع تلك اللغات. في الفصل التالي، سوف نعرض مهمة التعرف على كيانات الأسماء وسنبين كيف يمكن بناء أدوات المعالجة اللغوية التي ورد شرحها في هذا الفصل لإنجاز هذه المهمة.

الفصل الثالث
التعرف على كيانات الأسماء وتصنيفها

٣-١ مقدمة

كما ناقشنا في الفصل الأول، استخراج المعلومات هي عملية استخلاص المعلومات من النصوص غير المنظمة وتحويلها إلى بيانات منظمة. تلعب مهمة التعرف على كيانات الأسماء وتصنيفها (NERC) دورًا محوريًّا هنا، حيث تشمل هذه المهمة التعرف على الأسماء الصحيحة في النصوص (مهمة التعرف على كيانات الأسماء واختصارها NER)، وتصنيفها إلى مجموعة من الفئات ذات الأهمية المحددة مسبقًا (مهمة تصنيف كيانات الأسماء واختصارها NEC). على عكس أدوات المعالجة المسبقة التي نوقشت في الفصل السابق، والتي تُعنى بالتحليل النحوي، تُعنى مهمة التعرف على كيانات الأسماء وتصنيفها (NERC) باستنباط الدلالات من المحتوى النصي تلقائيًّا. المجموعة الأساسية التقليدية لكيانات الأسماء، التي تم تطويرها لمهمة NERC المشتركة في مؤتمر (MUC-6)، تتضمن تعبيرات الأشخاص والمنظمات والمواقع والتواريخ والوقت، مثل باراك أوباما ومايكروسوفت ونيويورك و٤ تموز (يوليو) ٢٠١٥ وما إلى ذلك.

بشكل عام، تُعدُّ مهمة التعرف على كيانات الأسماء وتصنيفها (NERC) مهمة إضافة تعليقات وشروحات annotation، بمعنى إضافة حواشٍ على شكل كيانات أسماء (NEs) إلى نص معين، ولكن يمكن أن يقتصر عملها ببساطة على إنتاج قائمة تضم كيانات أسماء يمكن استخدامها بعد ذلك لأغراض أخرى، بما في ذلك إنشاء أو توسيع معاجم كيانات الأسماء للمساعدة في إنجاز مهمة إضافة حواشي كيانات الأسماء إلى النصوص في المستقبل. يمكن تقسيم هذه المهمة إلى مهمتين: مهمة التعرف على كيانات الأسماء، التي تشتمل على التعرف على حدود كيانات الأسماء، (يشار إليها عادة باسم مهمة التعرف على كيانات الأسماء NER)؛ ومهمة تصنيف كيانات الأسماء (NEC)، وتشتمل على الكشف عن فئة أو نوع كيانات الأسماء. تُستخدم مهمة التعرف على كيانات الأسماء في الغالب لتعني كلتا المهمتين، على الرغم من كون ذلك قد يسبب بعض الالتباس، خصوصًا في الأعمال القديمة. في هذا الكتاب، سوف نتقيّد باستخدام مهمة التعرف على كيانات الأسماء وتصنيفها (NERC) لتعني

كلتا المهمتين، ومهمة التعرف على كيانات الأسماء لتعني عنصر التعرف على كيانات الأسماء فقط. لكي تكون مهمة تصنيف كيانات الأسماء أكثر دقة من التصنيف المعتاد الذي يقسّم كيانات الأسماء إلى أشخاص ومنظمات ومواقع، تؤخذ فئات الكيانات عادة من مخطط أنطولوجيا، وتكون فئات فرعية لتلك التصنيفات المعتادة [26]. يتمثل التحدي الرئيس الذي يواجه مهمة تصنيف كيانات الأسماء (NEC) في أن كيانات الأسماء يمكن أن تكون على درجة عالية من الغموض (على سبيل المثال: «ماي May» يمكن أن يكون اسم شخص ما أو أحد أشهر السنة؛ كما يمكن أن يكون «مارك Mark» اسما لشخص ما أو اسمًا شائعا. ولهذا السبب جزئيًّا، تُنفذ مهمة التعرف على كيانات الأسماء ومهمة تصنيف كيانات الأسماء كمهمة واحدة في العادة).

هناك مهمة إضافية تتعلق بكيانات الأسماء، وهي مهمة ربط كيانات الأسماء (NEL). تحدد هذه المهمة ما إذا كانت الإشارة إلى أحد كيانات الأسماء التي ترد في نص معين متوافقة مع أيّ كيانٍ من كيانات الأسماء الواردة في قاعدة معرفية مرجعية. تعني الإشارة إلى أحد كيانات الأسماء تعبيرًا يرد في النص للإشارة إلى أحد كيانات الأسماء: قد يرد هذا التعبير بأشكال مختلفة، على سبيل المثال، «السيد سميث» و»جون سميث» كلتاهما إشارتان (تمثيلان نصيّان) لكيان واحد في العالم الحقيقي، ويعبران عنه بتحقيقين لغويين مختلفين قليلاً. تكون القاعدة المعرفية المرجعية المستخدمة عادة موسوعة ويكيبيديا. مهمة ربط كيانات الأسماء (NEL) أكثر صعوبة من مهمة تصنيف كيانات الأسماء (NEC)، لأن تحديد أوجه التمايز بين الكيانات لا ينبغي أن يتم على مستوى فئة الكيان فحسب، بل يجب أن يتم أيضًا داخل فئات الكيانات. عل سبيل المثال، هناك أشخاص كُثر يحملون اسم «جون سميث». كلما كانت الأسماء شائعة أكثر، كلما أصبحت مهمة ربط كيانات الأسماء أكثر صعوبة. هناك مشكلة إضافية تتعلق بجميع المهام ذات الصلة بالقواعد المعرفية، وهي مشكلة عدم اكتمال القواعد المعرفية. على سبيل المثال، تتضمن هذه القواعد الأشخاص الأكثر شهرة ممن يحملون اسم «جون سميث». غير أن الأمر يشكل تحديًا من نوع خاص عند التعامل مع المهام التي تشتمل على أحداث جرت في الآونة الأخيرة، لأنه عادة ما يكون هناك فارق زمني بين الكيانات الناشئة حديثًا التي تبرز في الأخبار أو في شبكات التواصل الاجتماعي،

وبين عملية إضافة معلومات هذه الكيانات إلى القواعد المعرفية لغرض تحديثها. في الفصل الخامس سنورد مزيدا من التفاصيل بشأن مهمة ربط كيانات الأسماء، إلى جانب المكانز المرجعية ذات الصلة.

٣-٢ أنواع كيانات الأسماء

يرجع السبب في ارتفاع شعبية كيانات من قبيل الأشخاص والمنظمات والمواقع والتواريخ والأوقات كأنواع قياسية لتقسيم كيانات الأسماء إلى حد كبير إلى سلسلة مؤتمرات فهم الرسائل (MUC) [25]، التي استحدثت مهمة التعرف على كيانات الأسماء وتصنيفها في عام 1995م، والتي كانت بدورها القوة الدافعة وراء تطوير العديد من الأنظمة التي لا تزال موجودة اليوم. وبسبب التوسع في الجهود المبذولة لتقييم مهمة التعرف على كيانات الأسماء وتصنيفها (سيرد شرحها بشكل مفصل في القسم 3-3) والحاجة إلى استخدام أدوات مهمة التعرف على كيانات الأسماء وتصنيفها في تطبيقات عملية في سيناريوهات حقيقية، باتت تُعرف أنواع أخرى من الأسماء الصحيحة والتعبيرات تدريجيًّا على أنها كيانات أسماء، بما في ذلك الصحف والمبالغ النقدية، بالإضافة إلى التصنيفات الأدق للكيانات المشار إليها أعلاه، مثل المؤلفين والفرق الموسيقية وفرق كرة القدم والبرامج التلفزيونية، وما إلى ذلك. تعد مهمة التعرف على كيانات الأسماء وتصنيفها نقطة الانطلاق للعديد من التطبيقات والمهام المعقدة، مثل بناء الأنطولوجيات واستخراج العلاقات والإجابة عن الأسئلة واستخراج المعلومات واسترجاع المعلومات والترجمة الآلية وإضافة التعليقات والشروحات الدلالية. مع ظهور سيناريوهات استخراج المعلومات المفتوحة التي تشمل شبكة الإنترنت بأكملها، وتحليل محتوى شبكات التواصل الاجتماعي التي تظهر فيها كيانات جديدة باستمرار، ومهام ربط كيانات الأسماء، فقد اتسع نطاق الكيانات المستخلصة بشكل كبير، الأمر الذي جلب العديد من التحديات الجديدة (انظر على سبيل المثال القسم ٤-٤ الذي يناقش دور قواعد المعرفة في مهمة ربط كيانات الأسماء). علاوة على ذلك، باتت مهمة التعرف على الكيانات المعتادة المكونة من ٥ أو ٧ فئات تصنيفية أقل فائدة في الغالب، وهذا بدوره يعني أن هناك حاجة لتطوير نماذج جديدة. في بعض الحالات، مثل التعرف على أسماء مستخدمي تويتر، أصبح التمييز بين فئات الكيانات التقليدية، مثل المنظمات

والمواقع، غير واضح حتى بالنسبة للإنسان، ولم يعد هذا النوع مفيدًا في جميع الحالات (انظر الفصل الثامن).

إن تعريف ما ينبغي أن يكون عليه كل نوع من أنواع الكيانات ليس أمرًا سهلا على الإطلاق، وتختلف القواعد الإرشادية في هذا الخصوص تبعًا للمهمة. من الناحية التقليدية، كان الناس يستخدمون القواعد الإرشادية المعيارية الصادرة من مؤتمرات التقييم، مثل مؤتمر فهم الرسائل (MUC) ومؤتمر تعلم اللغات الطبيعية (CONLL)، لأن هذه المبادئ تسمح بالمقارنة بين الأساليب والأدوات بسهولة. لكن مع بدء استخدام الأدوات في تطبيقات عملية في سيناريوهات حقيقية، ولذا فمع تغيّر أنواع كيانات الأسماء وتطورها، فقد أصبح من الضروري أيضًا تكييف طرق تعريف الكيانات لتتلاءم مع المهمة. بطبيعة الحال، هذا الأمر يجعل عملية إجراء المقارنات وتقييم الأداء في الوقت الحالي أكثر صعوبة. على وجه الخصوص، سعى تقييم ACE [27] إلى حل بعض المشكلات الناجمة عن عملية تبديل الكلمات أو الكناية، التي يتم فيها استخدام كيان معين يصف من الناحية النظرية نوعًا محددًا من أنواع الكيانات (على سبيل المثال: منظمة) على نحو مجازي. من الأمثلة على ذلك فرق كرة القدم، حيث يجوز استخدام مواقع من قبيل إنجلترا أو ليفربول للإشارة إلى فريقي هذين الموقعين (على سبيل المثال: فازت إنجلترا بكأس العالم في عام 1966). وبالمثل، يمكن استخدام مواقع مثل البيت الأبيض أو ١٠ داوننغ ستريت للإشارة إلى المنظمة أو الهيئة التي توجد بداخلها (أعلن البيت الأبيض تعهدات بشأن المناخ أقرها ٨١ بلدًا). تشمل القرارات الأخرى مثلاً تحديد ما إذا كان ينبغي إدراج الذات الإلهية والرسل ضمن فئة «شخص»، وإذا كان الأمر كذلك، يُضاف إلى ذلك تحديد ما إذا كان ينبغي إدراجها في تلك الفئة في جميع الحالات، بما فيها الحالات التي يُستخدم فيها اسم الذات الإلهية والرسل كجزء من الألفاظ النابية.

٣-٣ تقييم كيانات الأسماء والمكانز

كما ذُكر أعلاه، كانت سلسلة مؤتمرات فهم الرسائل (MUC) أول سلسلة مهمة في مؤتمرات تقييم مهمة التعرف على كيانات الأسماء وتصنيفها NERC، حيث تناولت هذه السلسلة أول مرة التحدي الذي تمثله كيانات الأسماء في عام ١٩٩٦م. كان الهدف من ذلك التعرف على كيانات الأسماء الواردة في النص الإخباري، وهو ما لم يسهم

في تطوير نظام جديد فحسب، بل أدى أيضًا أول مرة إلى إصدار مكانز تحتوي على تعليقات وشروحات مكونة من كيانات أسماء، لتصبح هذه المكانز بمنزلة المعيار الذهبي المستخدم لأغراض التدريب والاختبار. وأعقب ذلك سلسلة مؤتمرات تعلم اللغات الطبيعية (CONLL) [28] في عام ٢٠٠٣م، وهي سلسلة أخرى ضمن مؤتمرات التقييم الرئيسة، وقد أصدرت بدورها بيانات أصبحت بمنزلة المعيار الذهبي لوكالات الأنباء، ليس فقط باللغة الإنجليزية، ولكن أيضًا باللغات الأسبانية والهولندية والألمانية. يعد المكنز الصادر عن هذه المؤتمرات حاليًّا من أكثر المعايير الذهبية شعبية في مهام التعرف على كيانات الأسماء وتصنيفها، وعادة ما تعتمد إصدارات برامج التعرف على كيانات الأسماء وتصنيفها على هذا المكنز فيما يتعلق بالأداء.

بدورها بدأت مؤتمرات التقييم الأخرى التي عقدت في وقت لاحق تتناول مسألة استخدام مهمة التعرف على كيانات الأسماء وتصنيفها في أنواع أخرى من النصوص غير الإخبارية، خصوصًا مكنز ACE [27] ومكنزOntoNotes [29]، واستحدثت أنواعًا جديدة من كيانات الأسماء. كلا هذين المكنزين يحتويان على مكانز فرعية تضم أنواعًا مختلفة من النصوص مثل نصوص وكالات الأنباء والبث المباشر للأخبار والبث المباشر للمحادثات ومدونات الويب والمحادثات التليفونية. بالإضافة إلى ذلك، يحتوي مكنز ACE على مكانز فرعية تحتوي على مجموعات أخبار فرعية في شبكة Usenet، ولا يقتصر على اللغة الإنجليزية فحسب، بل شمل أيضًا اللغات العربية والصينية في الإصدارات اللاحقة. يتضمن كل من مكنز ACE ومكنزOntoNotes أيضًا مهام مثل إيجاد جميع التعبيرات التي تشير إلى الكيان نفسه في النص، واستخراج العلاقات والأحداث، وإزالة الغموض في معاني الكلمات، مما يسمح للباحثين بدراسة التفاعل بين هذه المهام. سوف نتناول هذه المهام في القسم ٣-٥ وفي الفصلين الرابع والخامس.

وعلى الرغم من أن مكانز مهام التعرف على كيانات الأسماء وتصنيفها تستخدم في الغالب الأنواع التقليدية للكيانات، مثل الأشخاص والمنظمات والمواقع، وهي أنواع لا تستند إلى قاعدة معرفية صلبة للويب الدلالي (مثل DBpedia أو Freebase أو YAGO)، ولذلك فإن هذه الأنواع التقليدية عامة جدًّا. يعني ذلك أنه عند تطوير منهجيات مهام التعرف على كيانات الأسماء وتصنيفها بناءً على هذه المكانز لأغراض

الويب الدلالي، من السهل نسبيًّا البناء عليها وتضمين روابط لإحدى القواعد المعرفية فيها في وقت لاحق. على سبيل المثال، تستخدم أنطولوجيا NERD [30] أنطولوجيا OWL[1] التي تحتوي على مجموعة من المخططات لجميع فئات الكيانات (على سبيل المثال: فئة مجرم هي فئة فرعية لفئة شخص في أنطولوجيا NERD).

٣-٤ تحديات التعرف على كيانات الأسماء

أحد التحديات الرئيسة التي تواجهها مهمة تمييز كيانات الأسماء وتصنيفها تكمن في التمييز بين كيانات الأسماء وبين الكيانات الأخرى. وجه الاختلاف بينهما يكمن في أن كيانات الأسماء هي نماذج لأنواع الكيانات (مثل: شخص، سياسي) ويكون الكيان الذي تشير إليه كيانًا فريدًا واحدًا يوجد في واقع الحياة، في حين أن الكيانات الأخرى غالبًا ما تكون مجموعات من كيانات الأسماء التي لا تشير إلى كيانات فريدة موجودة في العالم الحقيقي. على سبيل المثال، «رئيس الوزراء» هو كيان، لكنه ليس كيانًا لاسم، لأنه يشير إلى أي شخص ينتمي إلى مجموعة من كيانات الأسماء (أي شخص شغل منصب رئيس الوزراء سابقًا أو حاليًّا). ومن الجدير بالذكر أن التمييز بينهما يمكن أن يكون صعبًا جدًّا، حتى بالنسبة للإنسان، مع العلم أن قواعد إضافة التعليقات والشروحات للمهام تختلف فيما بينها في هذا الشأن.

هناك تحد آخر يتمثل في التعرف على حدود كيانات الأسماء بشكل صحيح. في المثال ٣-١، من المهم إدراك أن كلمة السيد هي جزء من الاسم السيد روبرت والبول. لاحظ أن المهام تختلف أيضًا في المكان الذي تضع فيه حدود كيانات الأسماء. تنص المبادئ التوجيهية لمؤتمرات فهم الرسائل على أنه ينبغي أن تتضمن كيانات الأشخاص الألقاب، لكن مؤتمرات التقييم الأخرى قد تحدد مهامها بشكل مختلف. في المرجع [31] مناقشة جيدة لمشكلات تصميم مهام التعرف على كيانات الأسماء وتصنيفها، والاختلافات القائمة بينها. تعريفات الكيانات وحدودها غير متسقة في كثير من الأحيان، وهذا يعتمد على المكانز المختلفة. في بعض الأحيان، يعدُّ التعرف على حدود الكيانات مهمة منفصلة عن مهمة تحديد نوع كيانات الأسماء (شخص، موقع، ...الخ). هناك العديد

1- http://nerd.eurecom.fr/ontolog

من صيغ إضافة التعليقات والشروحات التي تُستخدم عادة للتعرف على مكان بداية كيانات الأسماء ومكان نهايتها. من بين صيغ التعليقات والشروحات الأكثر شعبية صيغة BIO، حيث يشير حرف B إلى Beginning أي بداية كيان اسم، ويشير حرف I إلى Inside أي داخل كيان اسم، ويشير حرف O إلى أن Ouside، أي أن الكلمة هي مجرد كلمة عادية تقع خارج نطاق كيان الاسم. هناك صيغة أخرى من صيغ التعليقات والشروحات تحظى بشعبية كبيرة، وهي صيغة BILOU [32]، التي تحتوي على ملصقات تصنيف إضافية هي حرف L (يشير إلى كلمة Last، ويعني آخر كلمة في كيان الاسم) وحرف U (يشير إلى كلمة Unit، ويعني أن الكلمة هي وحدة كاملة، أي كيان اسم).

مثال ٣-١ كان السيد روبرت والبول رجل دولة بريطانيًّا يعدُّ عمومًا أول رئيس وزراء لبريطانيا العظمى. على الرغم من أن التواريخ الدقيقة لفترة حكمه هي محل نقاش علمي، لكن فترة رئاسته على الأرجح في الفترة من ١٧٢١ إلى ١٧٤٢.[1]

سياسي: المناصب الحكومية التي شغلها (المسؤول، المركز/ المنصب/ اللقب، من، إلى)

شخص: الجنس

السيد روبرت والبول: سياسي، شخص

المناصب الحكومية التي شغلها (السيد روبرت والبول، رئيس وزراء بريطانيا العظمى، ١٧٢١، ١٧٤٢)

الجنس (السيد روبرت والبول، ذكر)

يعد الغموض من أكبر التحديات الماثلة أمام نظم التعرف على كيانات الأسماء وتصنيفها. يمكن أن يؤثر ذلك على العنصرين كليهما في مهمة التعرف على كيانات الأسماء وتصنيفها، وهما عنصر التعرف وعنصر التصنيف، كما يؤثر أحيانًا على العنصرين كليهما في الوقت نفسه. على سبيل المثال، يمكن أن تكون كلمة May (مايو)

١ – المثال من http://en.wikipedia.org/wiki/Robert_Walpole

اسم علم (كيانًا لاسم) أو اسم نكرة (وليس كيانًا، كما هو الحال في صيغة الفعل you may go (يمكنك الذهاب))، ولكن حتى عندما تكون كلمة May اسما، فإنها يمكن أن تندرج تحت فئات مختلفة (أحد أشهر السنة، أو جزءًا من اسم شخص ما (وفي هذه الحالة قد تشير إلى اسم الشخص أو لقبه)، أو جزءًا من اسم إحدى المنظمات). تحدث مشكلات التصنيف بصورة متكررة عند التمييز بين شخص ومنظمة، حيث تحمل العديد من الشركات أسماء أشخاص (على سبيل المثال: شركة الملابس Austin Reed). وبالمثل، تحمل العديد من الأشياء التي قد لا تكون كيانات أسماء، مثل أسماء الأمراض والقوانين، أسماء أشخاص أيضًا. على الرغم من أنه يمكن للمرء من الناحية الفنية إضافة تعليقات وشروحات لاسم الشخص هنا، إلا أن ذلك ليس مرغوبًا فيه عادة (نحن لا نهتم في العادة بإضافة التعليقات والشروحات لكلمة باركنسون لتحديد أنها تشير إلى شخص عندما ترد مثلاً في مصطلح مرض باركنسون أو كلمة فيثاغورس في نظرية فيثاغورس).

٣-٥ المهام المترابطة

تحويل النص الزمني للشكل القياسي (Temporal normalization) هي مهمة التعرف على التعبيرات الزمنية (كيانات الأسماء المصنفة كتاريخ أو وقت) وذلك بتحويلها إلى الصيغة المعيارية للتواريخ والأوقات. يعد تحويل النص الزمني للشكل القياسي، ولا سيّما التحويل للتواريخ والأوقات النسبية، ضروريًّا لمهام التعرف على الأحداث. تكون المهمة في غاية السهولة إذا كان النص يشير أصلاً إلى الوقت بصيغة مجردة، على سبيل المثال «٨ صباحًا». وتصبح المهمة أصعب إذا كان النص يشير إلى الوقت بصيغة نسبية، على سبيل المثال «الأسبوع الماضي». في هذه الحالة، يتعين علينا أولاً تحديد وقت إنشاء النص، وذلك لاستخدامه كنقطة مرجعية للتعبير الزمني النسبي. يعدُّ نظام TimeML [33] من بين أشهَر أنظمة إضافة التعليقات والشروحات الخاصة بالتعبيرات الزمنية. لا تتضمن غالبية أدوات التعرف على كيانات الأسماء وتصنيفها تحويل النص الزمني للشكل القياسي كجزء متعارف عليه من عملية التعرف على كيانات الأسماء وتصنيفها، لكن بعض الأدوات تتضمن ملحقات إضافية يمكن استخدامها لهذا الغرض. على سبيل المثال، يوجد في نظام GATE ملحق لتحويل

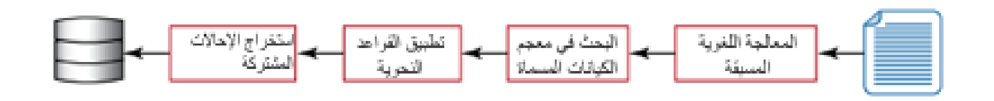
الوقت للشكل القياسي يمكن إضافته إلى نظام ANNIE. كما يتضمن ملحقًا لإضافة التعليقات والشروحات الزمنية، يسمى GATE–Time، وهو مبني على مُصنِّف HeidelTime [34] ويتوافق مع معيار TimeML، وهو معيار آيزو (ISO) خاصٌّ بالتعليقات والشروحات الزمنية الدلالية للوثائق [35]. SUTime [36] هي مكتبة أخرى للتعرف على التعبيرات الزمنية وتحويلها للشكل القياسي، وهي متوفرة كجزء من منظومة Stanford CoreNLP. تستخدم هذه المكتبة نظامًا حتميًّا يعتمد على القواعد، ومن ثَمَّ يمكن إضافة الملحقات إليها بسهولة. تُنتج هذه المكتبة مجموعة من التعليقات والشروحات التي تندرج تحت أحد الأنواع الزمنية الأربعة (تاريخ، وقت، مدة، مجموعة) المتوافقة مع معيار TIMEX3 الخاص بالنوع والقيمة. يشير النوع الزمني «مجموعة» غير المعتاد إلى مجموعة من الأوقات، مثل حدث متكرر.

استخراج الإحالات المشتركة (Co–reference resolution) يهدف إلى الربط بين الإشارات المختلفة للكيان نفسه. وتعد هذه المهمة ذات أهمية نظرًا لأنها تساعد في إيجاد العلاقات بين الكيانات في وقت لاحق، كما تساعد كذلك في الربط بين كيانات الأسماء. قد تكون الإشارات المختلفة إشارات متطابقة، وفي هذه الحالة تكون المهمة سهلة، وقد تكون المهمة أكثر تعقيدًا لأنه يمكن الإشارة إلى الكيان نفسه بطرق مختلفة. على سبيل المثال، جون سميث والسيد جون سميث وجون ج. س. سميث وسميث هي كلها إشارات إلى الشخص نفسه. وبالمثل، قد يكون لدينا اختصارات (.U.K وUnited Kingdom) أو حتى أسماء مستعارة لا تحمل وجه شبه بأسمائها البديلة من الناحية الخارجية آي بي إم وذا بيغ بلو (IBM وthe big blue). باستثناء الصيغة الأخيرة، التي يكون فيها الحل الأفضل استخدام قوائم مكونة من أسماء ثنائية صريحة، تميل الأنظمة المبنية على القواعد إلى تقديم أداء فعال في هذه المهمة. على سبيل المثال، على الرغم من كون الاختصارات شديدة الغموض في الغالب، لكن عندما يقتصر السياق الذي نتحدث عنه على الوثيقة نفسها أو المستند، نادرًا ما يحدث عدم تطابق بين اسم مختصر واسم كامل يتطابق مع الأحرف المعنية. بطبيعة الحال، يمكن أيضًا استخدام قوائم مكونة من أسماء ثنائية صريحة، كما يمكن كذلك إضافة قوائم الاستثناءات. تعدُّ أداة Orthomatcher الخاصة بمنصة ANNIE مثالاً جيدًا على الأدوات الخاصة بتحديد الإحالات المشتركة والتي تعتمد اعتمادًا كاملاً على القواعد المشفرة يدويًّا، حيث تعالج

هذه الأداة النصوص الإخبارية بدقة تصل إلى نحو ٩٥٪ [37]. أداة Stanford Coref مدمجة في منظومة Stanford CoreNLP، وتستخدم نظامًا متعدد التمرير لاستخراج الإحالات المشتركة والإحالات القبْلية وقد تم شرح النظام في المرجع [38]. يأتي نظام SANAPHOR بوظائف إضافية عن طريق إضافة طبقة دلالية إلى ما سبق وتحسين النتائج. تكون مدخلات هذا النظام عبارة عن مجموعات من الإحالات المشتركة يتم توليدها بواسطة أداة Stanford Coref، وبعد ذلك يقوم بفصل المجموعات التي تحتوي على إشارات غير مترابطة، بينما يدمج بين المجموعات التي ينبغي أن ينتمي بعضها إلى بعض. كما يستخدم مخرجات عمليات ربط كيانات الأسماء التي تُستخدم فيها قواعد المعرفة DBpedia وYAGO لإزالة الغموض عن الإشارات المتعلقة بكيانات مختلفة، ويدمج بين الإشارات المتعلقة بالكيان نفسه. يمكن استخدامه أيضًا في مهام التعرف على كيانات الأسماء وتصنيفها ومهام ربط كيانات الأسماء إلى جانب الأدوات الأخرى.

٣-٦ منهجيات التعرف على كيانات الأسماء وتصنيفها (NERC)

يمكن تقسيم منهجيات مهام التعرف على كيانات الأسماء وتصنيفها بشكل تقريبي إلى (١) منهجيات تستند إلى القواعد أو الأنماط، و(٢) أساليب التعلم الآلي أو الاستخراج الإحصائي [40]، وفي كثير من الأحيان يُمزج بين الأسلوبين (انظر [41][42][43]). تعتمد غالبية الأساليب القائمة على التعلم الآلي على شكل من أشكال الإشراف البشري، باستثناء أساليب استخراج المعلومات ذات الطبيعة الهيكلية البحتة التي تقوم بإجراء مهام التعلم الآلي غير الخاضعة للإشراف على مستندات تخلو من التعليقات والشروحات [44]. كما رأينا سابقًا، تتيح منصات هندسة اللغة مثل GATE وStanford CoreNLP وOpenNLP وNLTK تنفيذ أساليب وخوارزميات استخراج المعلومات على شكل وحدات، وذلك عن طريق إدراج وحدات معالجة مسبقة ووحدات خاصة بمهام التعرف على كيانات الأسماء وتصنيفها في منظومة التعرف على كيانات الأسماء، وهو ما يسمح بإخضاع نتائجها لتجارب وتقييمات قابلة للتكرار. يظهر الشكل ٣-١ مثالاً لمنظومة التعرف على كيانات الأسماء وتصنيفها.



الشكل ٣-١: منظومة التعرف على كيانات الأسماء وتصنيفها

٣-٦-١ المنهجيات القواعدية للتعرف على كيانات الأسماء وتصنيفها

الأساليب اللغوية المعتمدة على القواعد والمتعلقة بمهام التعرف على كيانات الأسماء، مثل الأساليب المستخدمة في نظام استخراج المعلومات ANNIE الخاص بمنصة GATE تتكون عادة من مزيج من معاجم كيانات الأسماء وقواعد مطابقة الأنماط المشفرة يدويًا. تستخدم هذه القواعد معلومات مأخوذة من السياق للمساعدة في تحديد ما إذا كانت الكيانات المحتملة الموجودة في معاجم كيانات الأسماء صحيحة، أو لزيادة عدد الكيانات المحتملة. تعدُّ معاجم كيانات الأسماء بمنزلة نقطة الانطلاق التي تتيح تأكيد أو رفض أو تنقيح الكيان النهائي الذي ينبغي استخراجه. تتكون منظومة التعرف على كيانات الأسماء وتصنيفها عادة من عملية معالجة لغوية مسبقة (تجزئة الجمل، تقسيم الجمل، تصنيف أقسام الكلام) كما سبق شرحه في الفصل السابق، تليها عملية إيجاد الكيان بواسطة معاجم كيانات الأسماء والقواعد النحوية، ثم عملية استخراج الإحالات المشتركة.

صُممت معاجم كيانات الأسماء لإضافة التعليقات والشروحات البسيطة والاعتيادية، مثل الأسماء المعروفة للشركات والمواقع وأيام الأسبوع والمشاهير وما إلى ذلك. قد تحتوي معاجم كيانات الأسماء النموذجية الخاصة بالتعرف على كيانات الأسماء وتصنيفها على مئات أو آلاف المدخلات. غير أن استخدام معاجم كيانات الأسماء ليس كافيًا بحد ذاته للتعرف على الكيانات وتصنيفها، وذلك لأن الكثير من الأسماء يتسم بالغموض (على سبيل المثال: «لندن» قد تكون جزءًا من اسم منظمة أو شخص، أو قد تكون المدينة المعروفة ببساطة) هذا من ناحية، ومن ناحية أخرى، لا يمكنها تحديد كل كيانٍ من كيانات الأسماء (على سبيل المثال: في اللغة الإنجليزية لا يمكن للمرء أن يحدد مسبقًا جنس كل لقب عائلي). لكن عند دمج معاجم كيانات الأسماء مع حواشي المعالجة اللغوية الأخرى (بطاقات تصنيف أقسام الكلام، الأحرف الكبيرة، وغيرها من الأدلة السياقية الأخرى)، فإنها قد تكون قوية جدًّا.

عملية مطابقة الأنماط في مهام التعرف على كيانات الأسماء وتصنيفها تتطلب تطوير الأنماط بناء على بنيات متعددة الجوانب تأخذ بعين الاعتبار العديد من الخصائص المختلفة للكلمات، بما فيها طريقة التهجئة (الكتابة بالأحرف الكبيرة في اللغة الإنجليزية) والإعراب والمعلومات الخاصة بتصنيف أقسام الكلام وما إلى ذلك. سرعان ما أصبحت عملية إدارة اللغات التقليدية المستخدمة في عمليات المطابقة بين الأنماط، كلغة PERL، شديدة الصعوبة بسبب التعقيد عند استخدامها في مهام من هذا القبيل. لذا عادة ما تُستخدم ترميزًا أو تدوينًا ثنائيًّا بصيغة «الخاصية- القيمة» والتي تسمح بأن تشير الشروط إلى خصائص بطاقات التصنيف الناجمة عن مستويات تحليل متعددة. من الأمثلة على ذلك لغة JAPE، وهي لغة لمطابقة الأنماط تعتمد على لغة جافا وتستخدم في نظام GATE، وهي مشتقة من لغة CPSL [45]. تستخدم لغة JAPE ترميزًا تعريفيًّا يسمح بكتابة قواعد قادرة على التعرف على السياق وإجراء عمليات مطابقة أنماط غير حتمية. تُقسّم القواعد إلى مراحل (مجموعات فرعية) يجرى تنفيذها بصورة متوازية، حيث تتكون كل مرحلة من المراحل عادة من قواعد خاصة بنفس نوع الكيان (على سبيل المثال: شخص) أو قواعد لها المتطلبات نفسها المحددة التي تكون شرطًا ضروريًّا لتنفيذها. تتيح مجموعة متنوعة من آليات تحديد الأولوية التعامل مع القواعد المتنافسة، وهو ما يجعل التعامل مع الغموض أمرًا ممكنًا: على سبيل المثال، قد يفضل المرء الأنماط التي تحدث في سياق معين، وقد يفضل نوعًا معينًا من أنواع الكيانات على نوع آخر في ظرف محدد. تعمل الآليات الأخرى المبنية على القواعد بطريقة مماثلة.

يمكننا تطبيق قاعدة نموذجية بسيطة لمطابقة الأنماط، قد تكون المهمة التي تقوم بها مطابقة جميع أسماء الجامعات، على سبيل المثال جامعة شيفيلد، جامعة بريستول. يتكون النمط من كلمة «جامعة» يليها اسم «المدينة». باستخدام معاجم كيانات الأسماء، يمكننا التحقق من ورود ذكر اسم مدينة ما مثل شيفيلد أو بريستول. أما القواعد الأكثر تعقيدًا، فيمكن استخدامها للتعرف على اسم أي منظمة من خلال البحث عن كلمة مفتاحية داخل معجم كيانات أسماء يرد ذكرها إلى جانب اسم علم واحد أو أكثر (حسبما تعثر عليه أداة تصنيف أقسام الكلام) مثل شركة، منظمة، مؤسسة تجارية، مدرسة، الخ، ويحتمل أيضًا أن تحتوي على بعض الكلمات الوظيفية. على الرغم من كون هذه الأنواع من القواعد فعالة جدًّا في مطابقة الأنماط المعتادة (ورغم كونها تعمل بشكل جيد مع

بعض أنواع الكيانات كالأشخاص والمواقع والتواريخ)، إلا أنها يمكن أن تكون شديدة الغموض. قارن مثلاً اسم الشركة General Motors (جنرال موتورز) واسم الشخص General Carpenter (الجنرال كاربنتر) وشبه الجملة Major Disaster (كارثة كبرى) (التي لا تشير إلى أي كيان)، لترى بسهولة أن مثل هذه الأنماط لا يؤدي الغرض بصورة كافية. على الجانب الآخر، قد يكون أداء المنهجيات التي تعتمد على التعلم جيدًا في التعرف على أن كلمة disaster (كارثة) لا تكون عادة جزءًا من اسم شخص أو منظمة، لأنها لا تظهر على هذا النحو مطلقًا في مكنز التدريب.

كما أوردنا سابقًا، يجري تطوير الأنظمة القواعدية بناءً على الخصائص اللغوية، مثل بطاقات تصنيف أقسام الكلام أو المعلومات المستقاة من السياق. وبدلاً من وضع هذه القواعد بصورة يدوية، من الممكن وضع علامات على الأمثلة التدريبية، ومن ثمّ تعلم القواعد بصورة آلية باستخدام أنظمة تعلم القواعد (تُعرف أيضًا بأنظمة استقراء أو استنتاج الأدلة). عن طريق التعلم الخاضع للإشراف، تقوم هذه الأنظمة باستنتاج مجموعات القواعد من الأمثلة التدريبية التي وُضعت عليها العلامات. كانت هذه الأنظمة تحظى بشعبية في أنظمة التعلم المبكرة التي كانت تُستخدم في مهام التعرف على كيانات الأسماء وتصنيفها، وكان من بينها أنظمة من قبيل SRV [46] وRAPIER [47] وWHISK [48] وBWI [49] وLP^2 [50].

٣-٦-٢ المنهجيات الخاضعة للإشراف للتعرف على كيانات الأسماء وتصنيفها

تاريخيًّا، ظهرت منهجيات التعلم الخاضع للإشراف بعد منهجيات التعلم المعتمدة على القواعد. تتعلم منهجيات التعلم الخاضع للإشراف أوزان الخصائص، وذلك بناءً على احتمال ظهورها في أمثلة تدريبية خاطئة مقابل أمثلة تدريبية صحيحة، وذلك لكل نوع محدد من أنواع كيانات الأسماء. بشكل عام، يتكون منهج التعلم الخاضع للإشراف من خمس مراحل:

- المعالجة اللغوية المسبقة؛
- استخراج الخصائص؛
- تدريب النماذج باستخدام البيانات التدريبية؛

- تطبيق النماذج على بيانات الاختبار؛
- المعالجة اللاحقة للنتائج لتصنيف المستندات.

المعالجة اللغوية المسبقة تشمل كحد أدنى تجزئة الجمل إلى وحدات لغوية وتقسيم الجمل. كما يمكن أن تشمل التحليل الصرفي وتصنيف أقسام الكلام واستخراج الإحالات المشتركة والتحليل الإعرابي، كما سبق شرحه في الفصل الثاني، وهذا يعتمد على الخصائص المستخدمة. تشمل الخصائص الشائعة ما يلي:

- الخصائص الصرفية: استخدام الأحرف الكبيرة [في اللغة الإنجليزية]، وجود الرموز الخاصة (مثال: $، ٪)؛
- خصائص أقسام الكلام: علامات ظهور كل قسم منها؛
- خصائص السياق: الكلمات الموجودة بجوار الكلمة المعنية وتصنيف أقسام الكلام التي تنتمي إليها هذه الكلمات، والتي تتراوح عادة بين كلمة واحدة وثلاث كلمات؛
- خصائص معجم كيانات الأسماء: ورود الكلمة المعنية في معاجم كيانات الأسماء؛
- الخصائص النحوية: خصائص مبنية على نتائج التحليل الإعرابي للجملة؛
- خصائص تمثيل الكلمات: الخصائص المبنية على التدريب غير الخاضع للإشراف باستخدام نص يخلو من ملصقات أو بطاقات التصنيف، على سبيل المثال: باستخدام طريقة براون لتجميع الكلمات (Brown clustering) أو تضمينات الكلمات (word embeddings).

تستخدم الأساليب الإحصائية للتعرف على الكيانات المسماة وتصنيفها تشكيلة متنوعة من النماذج، مثل نماذج ماركوف المخفية (HMMs) [51]، أو نماذج الإنتروبيا القصوى (Maximum Entropy models) [52]، أو آلات المتجه الداعم (SVMs) [53] [54] [55]، أو نماذج البيرسبترونز (Perceptrons) [56] [57]، أو الحقول الشرطية العشوائية (CRFs)[59, 58]، أو الشبكات العصبية [60]. المنهجيات الأكثر

نجاحًا في التعرف على كيانات الأسماء وتصنيفها تشمل المنهجيات المبنية على الحقول الشرطية العشوائية، والشبكات العصبية ذات المستويات المتعددة التي ظهرت حديثًا. وللمهتم بمعرفة المزيد عن خوارزميات التعلم الآلي يمكن الرجوع إلى [62, 61].

الحقول العشوائية الشرطية (CRF) تقوم بنمذجة مهمة التعرف على كيانات الأسماء وتصنيفها (NERC) لتكون بمنزلة منهجية للتصنيف بناء على متسلسلات، أي جعل بطاقات تصنيف الوحدات اللغوية يعتمد على بطاقات تصنيف الوحدات السابقة واللاحقة في جزء معين من التسلسل. من أمثلة أطر العمل المتاحة لمهام التعرف على كيانات الأسماء وتصنيفها (NERC) المبنية على الحقول الشرطية العشوائية إطار عمل Stanford NER[١] وإطار عمل CRFSuite[٢]. يحتوي كلاهما على أدوات لاستخراج الخصائص ونماذج مدربة باستخدام بيانات مؤتمر تعلم اللغات الطبيعية في عام ٢٠٠٣ (ConLL 2003) [28].

تتميز منهجيات الشبكات العصبية ذات المستويات المتعددة بميزتين. أولاً، تتعلم هذه المنهجيات الخصائص الكامنة أو الضمنية، بمعنى أنها لا تتطلب إجراء معالجة لغوية تتعدى تقسيم الجمل وتجزئتها إلى وحدات لغوية. هذا الأمر يجعلها أكثر فعالية في شتى المجالات مقارنة بالهياكل المبنية على الخصائص الصريحة، وذلك لأنها ليست مضطرة للتعويض عن الأخطاء التي تحدث أثناء إجراء المعالجة اللغوية المسبقة. ثانيًا، يمكنها أن تدمج بسهولة بين النصوص التي تخلو من العلامات التصنيفية، والتي يمكن تدريب أساليب استخراج الخصائص على تمثيلاتها. يستخدم نظام SENNA [60] المتطور الخاص بالتعرف على كيانات الأسماء وتصنيفها هيكلاً متعدد المستويات من الشبكات العصبية، إلى جانب تدريب غير خاضع للإشراف. يتوفر هذا النظام إما بشكل منفصل[٣] أو كجزء من إطار عمل DeepNL[٤]. ومثلما هو الحال مع أطر العمل المذكورة أعلاه، يتم توزيع هذا النظام مرفقًا بأدوات لاستخراج الخصائص، كما يوفر خاصية تدريب النماذج على بيانات جديدة.

1-http://nlp.stanford.edu/software/CRF-NER.shtml - http://www.chokkan.org/software/crfsuite/

2- http://ronan.collobert.com/senna/

3- https://github.com/attardi/deepnl

4- http://uima.apache.org

هناك مزايا وعيوب في منهجيات التعلم الخاضعة للإشراف عندما يتعلق الأمر بالتعرف على كيانات الأسماء وتصنيفها، مقارنة باستخدام منهجيات الهندسة المعرفية القواعدية. تتطلب كلتا المنهجيتين بذل جهد يدوي، إذ تتطلب المنهجيات القواعدية متخصصين لغويين ليقوموا بوضع قواعد مشفرة يدويًّا، في حين تتطلب المنهجيات القائمة على التعلم الخاضعة للإشراف بيانات تدريبية مشروحة، وهو ما يلغي الحاجة لوجود متخصصين لغويين. تعتمد المنهجية الأنسب لسيناريو تطبيقي معين على طبيعة التطبيق وعلى المجال. عندما يتعلق الأمر بالمجالات الشائعة، كالنصوص الإخبارية، تتوفر بيانات تدريبية مصنفة يدويًّا، في حين قد يكون من المطلوب إنشاء مثل هذه البيانات التدريبية بدءًا من الصفر بالنسبة للمجالات الأخرى. إذا كان التباين اللغوي في النص طفيفًا جدًّا، وهناك حاجة للحصول على النتائج بسرعة، فقد تكون القواعد المشفرة يدويًّا نقطة انطلاق أفضل.

٣-٧ أدوات التعرف على كيانات الأسماء وتصنيفها

يعد نظام ANNIE متعدد الأغراض الخاص بمنصة GATE المستخدم للتعرف على كيانات الأسماء وتصنيفها مثالاً نموذجيًّا للأنظمة القواعدية. صُمم هذا النظام لغرض التعرف على كيانات الأسماء وتصنيفها في النصوص الإخبارية، لكن نظرًا لسهولة تكييفه، يمكن أن يشكل نقطة الانطلاق للتطبيقات الجديدة في مجال التعرف على كيانات الأسماء التي يتم تطويرها للغات والمجالات الأخرى وتصنيفها. تتضمن منصة GATE أدوات للتعلم الآلي، ما يعني أنه يمكن استخدامها لتدريب نماذج التعرف على كيانات الأسماء وتصنيفها أيضًا، بناءً على مكونات المعالجة اللغوية المسبقة التي ورد شرحها في الفصل الثاني. تشمل الأنظمة الأخرى الأقل شهرة نظام UIMA[1]، المطور من قبل شركة آي بي إم، والذي يركز أكثر على الدعم الهيكلي وسرعة المعالجة، ويوفر عددًا من الموارد المماثلة لمنصة GATE، ونظام OpenCalaisK[2]، الذي يوفر خدمة ويب لتحشية النصوص بالدلالات لأنواع كيانات الأسماء التقليدية، ونظام LingPipe[3] الذي يقدم

1- http://www.opencalais.com/

2- http://alias-i.com/lingpipe/index.html

3- https://github.com/xiaoling/figer

مجموعة (محدودة) من نماذج التعلم الآلي لشتى المهام والمجالات. على الرغم من كون هذه الأنظمة عالية الدقة، إلا أنها ليست سهلة التكييف مع تطبيقات عملية جديدة. في واقع الأمر، توجد مكونات من جميع هذه الأدوات في نظام GATE، وذلك بهدف تمكين المستخدم من الجمع والتوليف بين الموارد المختلفة حسب الحاجة، أو المقارنة بين عمل الخوارزميات المختلفة على المكنز نفسه. غير أن المكونات المقدمة تكون بشكل عام على شكل نماذج سبق تدريبها، ولا توفر عادة جميع وظائف الأدوات الأصلية.

يعد نظام Stanford NER المرفق بمنظومة Stanford CoreNLP عبارة عن وحدة برمجية مكتوبة بلغة جافا للتعرف على كيانات الأسماء. يشتمل هذا النظام على أدوات ذات تصميم هندسي جيد للتعرف على كيانات الأسماء وتصنيفها، كما يوجد فيه عدد من الخيارات لتحديد هذه الأدوات. إضافة إلى النموذج المعتاد لكيانات الأسماء المكون من 3 فئات (الأشخاص، المنظمات، المواقع)، يتضمن هذا النظام أيضًا نماذج أخرى للغات المختلفة، ونماذج مدربة على مجموعات مختلفة. المنهجية التي يستخدمها هذا النظام هي تطبيق عام لنماذج تسلسلات الحقول الشرطية العشوائية ذات السلسلة الخطية، ولذا يمكن للمستخدم إعادة تدريبها بسهولة باستخدام أي بيانات مصنفة أخرى. يُستخدم نظام Stanford NER كذلك في منصة NLTK التي لا تتضمن أداة خاصة بها للتعرف على كيانات الأسماء وتصنيفها.

تحتوي منصة OpenNLP على وحدة NameFinder الخاصة بمهمة التعرف على كيانات الأسماء وتصنيفها (NERC) باللغة الإنجليزية، وبدورها تشتمل مهمة NERC على وحدات منفصلة خاصة بأنواع كيانات الأسماء السبعة المتعارف عليها وفقًا لتصنيف مؤتمرات فهم الرسائل (MUC) (شخص، منظمة، موقع، تاريخ، وقت، مال، نسبة مئوية)، وهي مدربة على قواعد بيانات قياسية متاحة مجانًا. تحتوي أيضًا على نماذج خاصة باللغتين الأسبانية والهولندية، وهي مدربة على بيانات مؤتمر تعلم اللغات الطبيعية (CONLL). وكما هو الحال مع أداة Stanford NER، بإمكان المستخدم إعادة تدريب وحدة NameFinder باستخدام أي بيانات مصنفة. وعلى غرار الأدوات الأخرى القائمة على التعلم المذكورة أعلاه، ونظرًا لاعتمادها على التعلم الخاضع للإشراف، تعمل هذه الأدوات بشكل جيد فقط عند وجود كميات كبيرة من

البيانات التدريبية المشتملة على الحواشي، لذا قد تكون هناك إشكالية عند تطبيقها على مجالات وأنواع نصوص جديدة إن لم توجد مثل هذه البيانات.

يعد نظام [63] FIGER[1] مثالًا للأنظمة التي تقوم بمهام التعرف على كيانات الأسماء وتصنيفها في مستويات تفصيلية دقيقة (fine-grained)، نظام FIGER مدرب على موسوعة ويكيبيديا. تتألف بطاقات التصنيف في نظام FIGER من ١١٢ نوعًا، وهي مشتقة من قاعدة Freebase المعرفية عن طريق اختيار الأنواع الأكثر تكرارًا ودمج الأنواع الأكثر دقة. يتمثل الهدف في إجراء تصنيف متعدد الفئات ومتعدد التصنيفات، بمعنى أن كل سلسلة من سلاسل الكلمات تُعطَى فئة واحدة أو عدة فئات، وقد لا تُعطَى أي فئة. يجري إعداد البيانات التدريبية لنظام FIGER عبر استغلال النص غير المشفر للكيانات المذكورة في حواشي وتعليقات موسوعة ويكيبيديا، بمعنى أن كل سلسلة من الكلمات الموجودة في جملة معينة تُربط بمجموعة من أنواع الكيانات الموجودة في قاعدة Freebase المعرفية، وتُستخدم كبيانات تدريبية إيجابية (صحيحة) لتلك الأنواع. يتم تدريب النظام باستخدام عملية مكونة من خطوتين، أولاهما تدريب نموذج حقل شرطي عشوائي للتعرف على حدود كيانات الأسماء، وثانيتهما تدريب خوارزمية بيرسبترون معدلة لتصنيف كيانات الأسماء. في العادة، يُستخدم نموذج حقل شرطي عشوائي للقيام بكلتا المهمتين في وقت واحد (مثال [64])، لكن يتم تجنب ذلك هنا بسبب المجموعة الكبيرة من أنواع كيانات الأسماء. وبخصوص الأدوات الأخرى للتعرف على كيانات الأسماء، يمكن إعادة تدريبها بسهولة باستخدام بيانات جديدة.

٣-٨ التعرف على كيانات الأسماء وتصنيفها في شبكات التواصل الاجتماعي

تعد الأبحاث في مجال التعرف على كيانات الأسماء في تغريدات تويتر وتصنيفها من مجالات البحث الساخنة، وذلك لوجود العديد من المهام التي تعتمد على تحليل محتوى شبكات التواصل الاجتماعي، كما سنناقش في الفصل الثامن. تمثل شبكات التواصل الاجتماعي تحديًا من نوعٍ خاص أمام مهام التعرف على كيانات الأسماء وتصنيفها، وذلك بسبب طبيعتها المشوشة (وجود أخطاء في الإملاء وعلامات الترقيم واستخدام

1- http://www.aclweb.org/aclwiki/index.php?title=CONLL-2003_(State_of_the_art)

الأحرف الكبيرة، واستخدام الكلمات بطرق مستحدثة، ...الخ)، وهو ما يؤثر في مكونات المعالجة المسبقة المطلوبة (ومن ثم يؤثر في أداء مكون التعرف على كيانات الأسماء وتصنيفها) وعلى كيانات الأسماء نفسها التي يصبح التعرف عليها أكثر صعوبة. ونظرًا لعدم وجود مكانز ذات حواشٍ وتعليقات، عادة ما يُنظر عمومًا إلى عملية التعرف على كيانات الأسماء في شبكات التواصل الاجتماعي وتصنيفها باستخدام منهجية تستند إلى التعلم على أنها مشكلة تتعلق بتكييف مهمة التعرف على كيانات الأسماء وتصنيفها مع مجالٍ جديد انتقالاً من النصوص الإخبارية، وغالبًا ما تدمج هذه العملية بين نوعي البيانات كليهما لغرض إجراء التدريب [65] وتتضمن خطوة إضافية وهي تحويل نص التغريدات إلى الشكل القياسي [66]. من بين التحديات المحددة تحدي المصطلحات (الكيانات) الحديثة، فغالبًا ما تكون أنواع كيانات الأسماء التي نريد التعرف عليها في شبكات التواصل الاجتماعي ناشئة حديثًا (على سبيل المثال قصص إخبارية حديثة تتعلق بأشخاص لم يكونوا مشهورين سابقًا) ولهذا لا تكون هذه الكيانات في العادة موجودة في معاجم كيانات الأسماء أو حتى في قواعد البيانات المترابطة مثل DBpedia. هناك تحدٍّ آخر وهو أن السياق المتنوع [67] وكذلك إطار السياق الأصغر [68] يجعل من الصعب التعرف على كيانات الأسماء وتصنيفها، فعلى عكس المقالات الإخبارية الطويلة، تتوفر كمية قليلة من معلومات الخطاب في كل تغريدة، والهيكل المتسلسل مجزأ عبر وثائق متعددة، كما يتدفق في اتجاهات متعددة. سنناقش عملية التعرف على كيانات الأسماء وتصنيفها في شبكات التواصل الاجتماعي بوضوح في الفصل الثامن.

٣-٩ الأداء

بشكل عام، يقل أداء مهمة التعرف على كيانات الأسماء وتصنيفها عن أداء مهام المعالجة المسبقة الموجودة في منظومة معالجة اللغات الطبيعية، مثل مهمة تصنيف أقسام الكلام، لكن يمكنه الوصول إلى درجات F1 تزيد نسبتها على ٩٠٪. يعتمد أداء مهمة التعرف على كيانات الأسماء وتصنيفها على مجموعة متنوعة من العوامل، بما فيها نوع النص (على سبيل المثال: النصوص الإخبارية، محتوى شبكات التواصل الاجتماعي) ونوع الكيان المسمى (مثال: شخص، موقع، منظمة) وحجم المكنز التدريبي المتوفر، والعامل الأهم هو مدى اختلاف المكنز الذي جرى على أساسه تطوير مهمة التعرف

على كيانات الأسماء عن النص الذي تُعالجه هذه المهمة [69]. في مؤتمرات المنافسة لتقييم عملية التعرف على كيانات الأسماء وتصنيفها، تتمثل المهمة عادة في تدريب الأنظمة واختبارها على أقسام مختلفة من المكنز نفسه (تُعرف أيضًا بالأداء داخل المجال)، بمعنى أن مكنز الاختبار يكون مشابهًا جدًّا لمكنز التدريب.

لإعطاء مؤشر على الأداء داخل المجال المشار إليه، يصل أداء النتائج الحديثة في مكنز مؤتمر تعلم اللغات الطبيعية لعام ٢٠٠٣ (ConLL 2003) الذي يعدُّ أشهر مكنز إخباري يتضمن حواشيَ وتعليقات التعرف على كيانات الأسماء وتصنيفها إلى F1 ١٠ , ٩٠٪. في الوقت الحالي، النظام الأفضل من حيث الأداء هو [1][70]. في المقابل، لم تحقق الأداة الفائزة بمهمة التعرف على كيانات الأسماء في شبكات التواصل الاجتماعي وتصنيفها خلال ورشة عمل المهام المشتركة حول المحتوى المشوش المُنتج على يد المستخدم لعام 2015 (WNUT) [71, 70] سوى نسبة أداء F1 ٤١ , ٥٦٪، كما حققت نسبة أداء ٦٣ , ٧٠ في مهمة التعرف على كيانات الأسماء. من الواضح أن مهمة التعرف على كيانات الأسماء وتصنيفها أكثر صعوبة بكثير من مهمة التعرف على كيانات الأسماء، وأن مهمة التعرف على كيانات الأسماء وتصنيفها في مكانز شبكات التواصل الاجتماعي الموجودة حاليًّا أكثر صعوبة من مهمة التعرف على كيانات الأسماء وتصنيفها في مكانز المحتوى الإخباري. جدير بالذكر أن المكانز تختلف أيضًا في حجمها، وهذا الأمر طبيعي. توجد مكانز ذات حواشٍ وتعليقات خاصة بالتعرف على كيانات الأسماء وتصنيفها للنصوص الإخبارية، إلا أنَّ محتوى شبكات التواصل الاجتماعي لا يزال يفتقر إلى حد كبير لمثل هذه المكانز. يشكل هذا الأمر سببًا مهمًّا من أسباب كون الأداء في مكانز شبكات التواصل الاجتماعي أسوأ بكثير [69]. ينطبق هذا الأمر بشكل خاص على محتوى شبكات التواصل الاجتماعي، حيث تتغير الكيانات بسرعة كبيرة. في الممارسة العملية، نعني بذلك أنه بعد بضع سنوات، قد تصبح بيانات التدريب المستخدمة الآن عديمة الجدوى تقريبًا.

1- http://www.nist.gov/tac/2014/KBP/SFValidation/index.html

٣-١٠ خلاصة

في هذا الفصل، شرحنا مهمة التعرف على كيانات الأسماء وتصنيفها والمهمتين الفرعيتين اللتين تشتمل عليهما، وهما مهمة التعرف على حدود الكيانات ومهمة تصنيف الكيانات إلى أنواع. كما أوضحنا سبب الحاجة إلى وجود التقنيات اللغوية التي ورد شرحها في الفصل السابق لإتمام هذه المهمة وكيفية استخدام تلك التقنيات في كل من منهجي التعلم القائم على القواعد والتعلم الآلي. وعلى غرار معظم مهام معالجة اللغات الطبيعية التالية التي سنشرحها في بقية الكتاب، تعد مهمة التعرف على كيانات الأسماء وتصنيفها النقطة التي تبدأ الصعوبة عندها بحيث تصبح المهام التالية أكثر تعقيدًا. بشكل أساسي، جميع المهام اللغوية التي تقوم بعملية المعالجة المسبقة لها هدف وتعريف متماثل جدًّا، وهذا الأمر لا يختلف تبعًا للغرض الذي ستُستخدم هذه المهام من أجله. تختلف مهمة التعرف على كيانات الأسماء، وكذلك المهام الأخرى من قبيل استخراج العلاقات وتحليل المشاعر وغيرهما، تختلف اختلافًا كبيرًا في تعريفاتها، وهذا يعتمد على سبب الحاجة لهذه المهام. على سبيل المثال، قد تختلف أنواع كيانات الأسماء اختلافًا شاسعًا عن أنواع الكيانات القياسية المعتمدة من قبل مؤتمرات فهم الرسائل (MUC)، وهي الأشخاص والمنظمات والمواقع، لتصبح أنواع كيانات الأسماء أكثر تفصيلاً ودقة وتشمل أنواعًا أكثر من ذلك بكثير، وهو ما يجعل طبيعة المهمة مختلفة جدًّا. من هنا يمكن للمرء أيضًا الذهاب خطوة أبعد وإضافة حواشٍ وتعليقات أكثر دلالة، وذلك عبر ربط الكيانات بمصادر بيانات خارجية مثل DBpedia وFreebase، كما سنرى في الفصل الخامس. على الرغم من ذلك، تتسم أساليب التعرف على كيانات الأسماء وتصنيفها بقابليتها للاستخدام المتكرر (في بعض السياقات) حتى عندما تختلف المهمة بصورة جوهرية، على الرغم من أن بعض أساليب التعلم الآلي مثلاً قد تعمل بطريقة أسوأ أو أفضل حسب مستويات تصنيف أنواع الكيانات المختلفة. في الفصل التالي، سوف نلقي نظرة على كيفية الربط بين كيانات الأسماء بواسطة العلاقات، مثل المؤلفين وكتبهم، أو الموظفين وشركاتهم.

الفصل الرابع
استخراج العلاقات

٤-١ مقدمة

تعنى مهمة استخراج العلاقات (RE) باستخراج الروابط بين العلاقات، وهذه المهمة تعتمد على مهمة التعرف على كيانات الأسماء التي ناقشناها في الفصل السابق. في العادة يكون محور الاهتمام في هذه المهمة استخراج العلاقات الثنائية بين كيانات الأسماء، لكنها قد تشمل أيضًا استخراج علاقات أكثر تعقيدًا مثل الأحداث. تشمل أنواع العلاقات عادة علاقات مثل تاريخ ميلاد (شخص، تاريخ) ومؤسس (شخص، منظمة)، وتشمل أمثلة العلاقات تاريخ ميلاد (جون سميث، ١٩٨٥-٠١-٠١) أو مؤسس كيان (بيل جيتس، مايكروسوفت).

قد تكون مهمة استخراج العلاقات مرتبطة بالتعليقات والشروحات، أي إضافة العلاقات والشروحات إلى النص، لكنها تعدُّ في العادة مهمة لملء الفتحات، كما تسمى أيضًا مهمة تعبئة قواعد المعرفة، أي تعبئة قاعدة معرفة معينة بالعلاقات لمجموعة معينة من أنواع العلاقة (تُعرف باسم مخطط العلاقة). يمكن تقسيم هذه المهمة إلى ثلاث مهام فرعية: تحديد معطيات العلاقة (إيجاد حدود المعطيات)، تصنيف معطيات العلاقة (تحديد أنواع المعطيات)، وتصنيف العلاقة (تحديد نوع العلاقة) [73]. بصفة عامة، يجري تنفيذ المهمتين الأوليين باستخدام عملية التعرف على كيانات الأسماء وتصنيفها. لإجراء عملية إضافة التعليقات والشروحات الدلالية (راجع القسم الخامس من هذا الفصل)، هناك خطوة إضافية تتمثل في ربط معطيات العلاقات بمدخلات قاعدة بيانات معينة باستخدام أساليب ربط كيانات الأسماء (NEL).

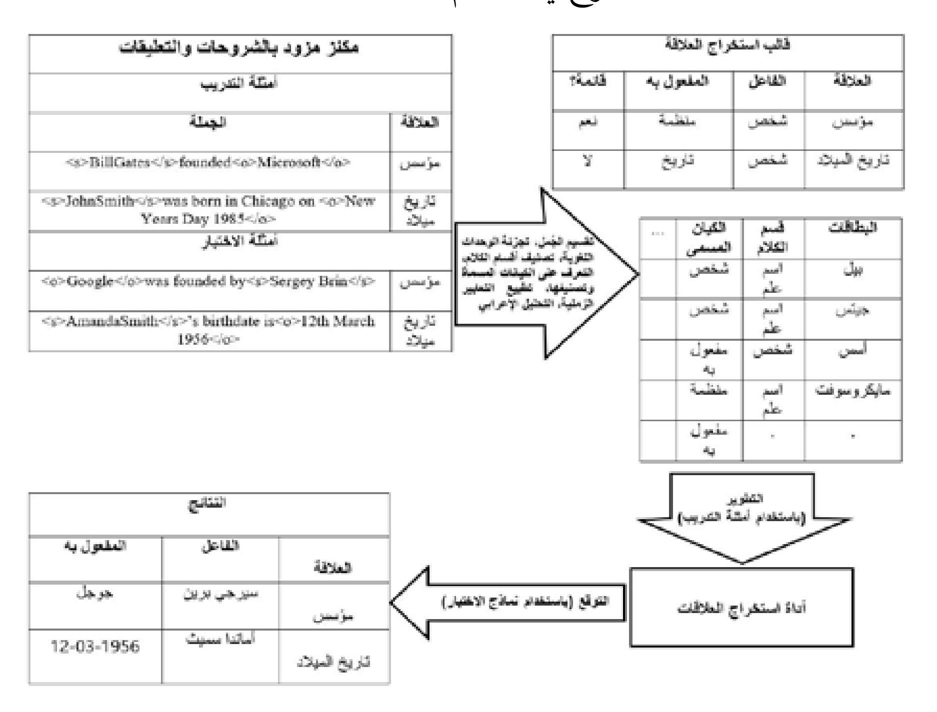
من بين المشكلات التي تواجهها منهجيات استخراج العلاقات الاختلاف الكبير بين مخططات العلاقات، فعلى عكس مهمة التعرف على كيانات الأسماء، لا توجد مجموعة صغيرة من أنواع الكيانات المعيارية مشتركة بين الأنظمة المختلفة. يعتمد المخطط المستخدم إلى حد بعيد على طبيعة التطبيق. في بعض الحالات، يُستخدم مخطط أنطولوجيا موجود حاليًّا، على سبيل المثال مخطط YAGO، في حين يجري إنشاء مخطط خاص بالمهمة في الحالات الأخرى. لهذا السبب، يقل عدد أنظمة استخراج العلاقات الجاهزة عن عدد أنظمة التعرف على كيانات الأسماء الجاهزة.

هناك مشكلة أخرى، وهي أن أنواع العلاقات قد تتداخل أو تتبع إحداها الأخرى، على سبيل المثال، الرئيس التنفيذي لـ(شخص، مؤسسة) هي علاقة تندرج بشكل كامل تحت علاقة موظف في (شخص، مؤسسة)، بينما يوجد تداخل قوي فقط، في المقابل لا توجد علاقة تستلزم التداخل بين الكيانين بلد الميلاد (شخص، موقع) وبلد الإقامة (شخص، موقع). في بعض الأحيان، يُحدد مخطط العلاقات الضمني تعريف علاقات التلازم هذه، ومن ثمَّ يمكن استخدامه لتحسين أداء عملية استخراج العلاقة [74].

أخيرًا، تجدر الإشارة إلى أنه كلما كانت العلاقة أشمل وأكثر تكرارًا، كان من السهل تحقيق أداء أعلى في عملية استخراج تلك العلاقة.

٤-٢ مسار عملية استخراج العلاقات

يهدف هذا القسم إلى تقديم شرح لمنهجية استخراج العلاقة النموذجية. يظهر الشكل ٤-١ نظرة عامة رسومية لمنظومة استخراج العلاقات. لاحظ أن هناك عدة أشكال لهذه المنهجية، كما سنشرح في الأقسام التالية.



الشكل ٤-١: مسار عملية استخراج العلاقات النموذجية

في العادة، تكون مُدخلات مهمة استخراج العلاقة عبارة عن مجوعة من الوثائق التدريبية ووثائق الاختبار وقالب استخراج العلاقة. يحدد قالب الاستخراج تعريف العلاقات التي ينبغي استخراجها وطريقة تعريفها، أي كم عدد المعطيات التي توجد فيها وما المفاهيم التي تنتمي إليها تلك المعطيات. على سبيل المثال، تُعرّف العلاقة (مؤسس) كعلاقة بين شخص (PER) ومنظمة (ORG): مؤسس (شخص، مؤسس)، وهي من علاقات القوائم، أي يحتمل أنها قد تتضمن أكثر من مفعول به واحد (مؤسس) لكل فاعل وعلاقة. لا تُعطى أنواع كيانات الأسماء بصورة مفصلة دائمًا، على سبيل المثال، لم تُقدم مهمة ملء الفتحات في مؤتمر تحليل النصوص لعام 2014 (TAC KBP 2014) نوع الكيان المسمى الخاص بالمفعول به في العلاقة [75]. تمر الوثائق بعد ذلك بعملية المعالجة المسبقة التي تشمل تنفيذ عدة خطوات تندرج ضمن عملية معالجة اللغة الطبيعية بهدف تحديد الطبيعة الصرفية والنحوية والدلالية للجُملة. تهدف خطوات المعالجة المسبقة هذه إلى المساعدة في «فهم» النص من أجل تسهيل عملية استخراج العلاقات.

تعدُّ مهمة التعرف على كيانات الأسماء وتصنيفها من أهم خطوات المعالجة اللغوية المسبقة، والسبب هو أن العلاقات تُستخلص إما بين كيانات الأسماء فقط، أو بين خليط من كيانات الأسماء والمفاهيم العامة (مثال: شخص)، كما ذكرنا في القسم السابق. على سبيل المثال، يُعطى الكيان بيل جيتس النوع شخص (PER) ويُعطى الكيان مايكروسوفت النوع مؤسسة (ORG). في الماضي، ميّزت الجهود الأولى التي بُذلت من أجل تقييم العلاقات خلال مؤتمرات فهم الرسائل بين أنواع كيانات الأسماء شخص (PER) وموقع (LOC) ومؤسسة (ORG) ومتفرقات (MISC) [25]، على الرغم من إمكانية استخدام أنواع مفصلة أكثر (مثل سياسي، فيلم)، وذلك حسب طبيعة قالب استخراج العلاقة.

بعد تنفيذ عملية المعالجة المسبقة، تُستخدم وثائق التدريب لتطوير أدوات استخراج العلاقات، وبعد ذلك تُطبق على وثائق الاختبار من أجل استخراج العلاقات. في حال استخراج أكثر من علاقة واحدة لكل قالب، يتم إثبات صحة تلك العلاقات المستخرجة. قد يكون تعريف العلاقات عاملاً مساعدًا في هذا الجانب. على سبيل

المثال، قد يكون لشركة ما أكثر من مؤسس واحد، لكن كل شخص لديه أبوان حقيقيان -ليس بالتَّبني- وبناءً على ذلك يتقرر عدد العلاقات التي ينبغي استخراجها لكل فاعل في كل علاقة.

تكون مخرجات عملية استخراج العلاقة على شكل مجموعة من وثائق الاختبار ذات حواشٍ (تُدعى غالبًا عملية استخراج على مستوى الجملة) أو على شكل قائمة مكونة من مستخلصات ثلاثية (استخراج على مستوى الكيان). في حال كون المخرجات على شكل قائمة مستخلصات، يمكن استخدامها لتعبئة قواعد المعرفة. يقدم القسم التالي مزيدًا من التفصيل عن قواعد المعرفة ودورها في مهمة استخراج العلاقات.

٤-٣ العلاقة بين مهمة استخراج العلاقات والمهام الأخرى

تُعرّف مهمة استخراج العلاقات بصفة عامة بأنها استخراج إشارات العلاقات إلى جانب معطياتها من النص. عند الحديث عن مهمة استخراج العلاقات التقليدية، تُعرّف أنواع العلاقات ومعطياتها داخل مخطط، في حين لا تُعرّف أنواع العلاقات مسبقًا عندما يتعلق الأمر بعملية استخراج المعلومات المفتوحة [76] حيث تكون أنواع العلاقات غير معرفة مسبقا. فعلى سبيل المثال (شخص، ولد في، تاريخ) هذا من الأمثلة على قالب العلاقات الثنائية، على الرغم من أن معطيات العلاقات قد تزيد عن اثنين، على سبيل المثال المناصب الحكومية. كما رأينا سابقًا، تبنى مهمة استخراج العلاقات على مهمة التعرف على كيانات الأسماء وتصنيفها، وذلك لأنه يجب تحديد الكيانات أولاً لكي تُستخلص العلاقات القائمة بينها.

هناك العديد من التحديات في مهمة استخراج العلاقات، فإلى جانب المشكلات الموجودة في عملية التعرف على كيانات الأسماء وتصنيفها، يتمثل التحدي الرئيس في إمكانية التعبير عن العلاقات بطرق مختلفة. على سبيل المثال، يمكن التعبير عن العلاقة (وُلِد) بعدة طرق، مثل (مولده في، أو تاريخ ميلاده في، أو أبصر النور للمرة الأولى في). إضافة إلى ذلك، ليست تعبيرات العلاقات خاصة بعلاقة واحدة دائمًا، على سبيل المثال، يمكن أن تعني العلاقة (يعمل في) إما (موظف في أو الرئيس التنفيذي لـ). تتسم بعض تعبيرات العلاقات أيضًا بالغموض الشديد، على سبيل المثال، عندما نقول: «الطيور»

لألفريد هيتشكوك كانت ذات شعبية واسعة. في تلك الحالة، يكون السياق مفيدًا جدًّا، أي بما أن ألفريد هيتشكوك كان صانع أفلام، من المرجح جدًّا أن الطيور كان فيلمًا. قد تمتد العلاقات أيضًا لتشمل عدة جُمل، وقد تحتوي فقط على إشارة غير مباشرة إلى أحد الكيانات المشمولة بالعلاقة (على سبيل المثال: الضمير: هُم)، كما يظهر في المثال التالي.

المثال ٤-١ في نوفمبر عام ١٩٦٣ وقّعت كابيتول ريكوردز عقدًا مع البيتلز وأعلنت عن خططٍ لإصدار الأغنية المنفردة «I Want To Hold Your Hand» (أريد أن أمسك بيدك) في شهر ديسمبر عام ١٩٦٣، إضافة إلى ألبومهم الثاني «With the Beatles» (مع البيتلز) في شهر يناير.

إذًا خطوات عملية المعالجة المسبقة مثل عملية استخراج الإحالات المشتركة تكون مفيدة. كما هو الحال مع كيانات الأسماء، يمكن إضافة التعليقات والحواشي إلى العلاقات الموجودة في النص، أو استخراجها واستخدامها لتعبئة قاعدة معرفة. لتعبئة قواعد المعرفة، هناك خطوة إضافية تتمثل في الدمج بين العلاقات المستخلصة، وتشكل هذه الخطوة أيضًا جزءًا من تحديات مؤتمر تحليل النصوص - تعبئة قواعد المعرفة (TAC KBP[1]). لدمج العلاقات المستخلصة، من المهم اتخاذ قرار بشأن ما إذا كانت العلاقات المستخلصة مترادفة، أو ما إذا كانت إحداها تتبع الأخرى، أو ما إذا كانت متناقضة. إذًا، فإن كُلاً من مهمة تمييز الالتزام النصي (recognizing textual entailment - RTE)، أي التعرف على إمكانية أن يُستنتج تعبير ما من تعبير آخر، ومهمة كشف التناقض (CD - contradiction detection)، أي استحالة أن تكون عبارتان صحيحتين في آن واحد، هاتان المهمتان مترابطتان مع أهميتهما كلتيهما.

مهمة استخراج الأحداث هي مهمة التعرف على الأحداث، والأحداث عبارة عن مجموعة من العلاقات التي غالبًا ما يكون لها مشاركون وتاريخ بداية وتاريخ نهاية وموقع. من الأمثلة على ذلك افتتاح مطعم. يجري افتتاح المطعم في نقطة معينة من الزمن، لكنه قد يُغلق ويُعاد فتحه مرة أخرى في موقع مختلف، ربما باسم مالك جديد. هناك صعوبة شديدة في عملية استخراج الأحداث، ويرجع السبب جزئيًّا إلى كون عملية الاستخراج تشمل التحليل الزمني، وبسبب الغموض الكبير في تعريف الحدث.

1- http://www.nist.gov/tac/2014/

على الرغم من أن تنفيذ عملية استخراج العلاقات يكون غالبًا على شكل مراحل متتالية، كما هو مبين في الشكل ٤-١، إلا أن ذلك قد يؤدي إلى انتقال الأخطاء من مرحلة إلى أخرى. ففي حال وقوع خطأ في مرحلة مبكرة من مراحل العملية، لا يمكن تصحيحه لاحقًا. على سبيل المثال، في حال فشل مهمة التعرف على كيانات الأسماء وتصنيفها في التعرف على كيان اسم، لن يكون بوسع أداة استخراج العلاقات تصحيح ذلك الخطأ. لهذا السبب، قد تُطرح حلول بديلة لهذه المسألة، حيث تتعلم هذه الحلول المهام المختلفة معًا. يسمح هذا الأمر باستخدام المعلومات الواردة في المراحل المتأخرة من عملية المعالجة (مثل مهمة استخراج العلاقات) وفي المراحل المبكرة (مثل مهمة التعرف على كيانات الأسماء وتصنيفها) من أجل تصحيح الأخطاء. تجدر الإشارة إلى أنه قد جرى طرح أساليب لمعالجة هذه المشكلة، حيث تقوم هذه الأساليب بتنفيذ مهمتي التعرف على كيانات الأسماء وتصنيفها واستخراج العلاقات معًا في آنٍ واحد [73, 77]، أو تنفيذ مهمة التعرف على كيانات الأسماء وتصنيفها ومهمة استخراج العلاقات ومهمة استخراج الإحالات المشتركة معًا في آنٍ واحد [78, 79].

٤-٤ دور قواعد المعرفة في استخراج العلاقات

تمثل قواعد المعرفة جزءًا أساسيًّا من عملية استخراج العلاقات. تتكون قواعد المعرفة من مخطط، ويُسمى هذا المخطط قالب استخراج في بعض الأحيان، بالإضافة إلى البيانات المرتبطة بالمخطط. يُعرّف المخطط هيكل المعلومات، على سبيل المثال، قد يُعرّف الأشخاص بأنهم سياسيون أو موسيقيون، وأن لهم أسماء وتواريخ ميلاد، وأن السياسيين يكونون مرتبطين بأحد الأحزاب بالإضافة إلى ما سبق، وأن الموسيقيين يعزفون على الآلات ضمن فِرقٍ موسيقية مع موسيقيين آخرين. إذًا، يُعرّف المخطط الفئات (مثال: شخص) وفئاتها الفرعية (مثال: سياسي) وخصائصها (مثال: داخل حزب). الجانب الذي يعني مهمة استخراج العلاقات هو أن الخصائص تحدد العلاقات التي يمكن أن تنشأ بين الفئات، في حين تقيّد فئاتها أنواع معطيات العلاقات. إذًا، تكون البيانات المرتبطة بالمخطط أمثلة على السياسيين والموسيقيين بأسمائهم وتواريخ ميلادهم وأحزابهم وآلاتهم الموسيقية وفِرقهم. تبدأ عملية استخراج العلاقات عادة بهذا المخطط، وبعد ذلك يصبح الهدف المنشود إضافة حواشٍ وتعليقات النص بالعلاقات، أو تعبئة

قاعدة المعرفة بالمعلومات، أي استخراج البيانات وإضافتها. تُعرف المهمة الأخيرة باسم تعبئة قاعدة المعرفة (KBP) وقد باتت تحظى بشعبية نظرا لسلسلة مؤتمرات تحليل النصوص - تعبئة قواعد المعرفة (TAC KBP) علاوة على وجود أسباب أخرى[1]. تتكون هذه السلسلة التي تُعنى بجهود التقييم من عدة أجزاء من مراحل منظومة استخراج العلاقات، بما في ذلك استخراج العلاقات (تعبئة الفتحات) [75] والتحقق من صحة العلاقات المستخلصة (التحقق من صحة معبئات الفتحات). في عملية تعبئة الفتحات، يكون الفاعل أو العلاقة جاهزة، وتتمثل المهمة بعد ذلك في إيجاد المفعول به في العلاقة داخل أحد المكانز.

غالبًا ما تستخدم جهود تقييم المهام المشتركة قوالب مُعرّفة محليًّا. غير أنه ومع بروز شبكة الإنترنت ومن بعدها الويب الدلالي، أصبحت قواعد المعرفة الموجودة على الإنترنت والمتاحة أمام الجمهور تحظى أيضًا بشعبية عندما يتعلق الأمر بمهمة تعبئة قواعد المعرفة [80, 81].

٤-٥ مخططات العلاقات

هناك نوعان من المعلومات التي ينبغي شرحها في عملية استخراج العلاقات. أولاً، نحن بحاجة إلى معلومات تتعلق بالفئات (على سبيل المثال: فنان، مقطوعة) والعلاقات التي تجمعها (على سبيل المثال: أصدر مقطوعة). يُنشر هذا النوع من المعلومات على شكل مخطط. ثانيًا، نحن بحاجة إلى معلومات عن الحالات المفردة لتلك الفئات (على سبيل المثال: ديفيد بوي، تغييرات Changes)، حيث يمكن نشر تلك المعلومات في قاعدة بيانات. لكن نلاحظ أن هذا الأمر اختياري: تحتوي بعض مواقع الإنترنت رموزًا دلالية تستخدم عادة http://schema.org/، لكنها لا تنشرها في قاعدة بيانات منفصلة.

على الرغم من أن المخططات تؤدي غرضًا مشابهًا لغرض القوالب المُعرّفة محليًّا (القسم ٤-٤) عندما يتعلق الأمر بمهمة استخراج العلاقات، إلا أن لها ميزة واضحة في طريقة وصف البيانات، حيث تُستخدم مُعرّفات مميزة للكيانات تسمى مُعرّفات الموارد الموحدة (URIs). تخيل مهمة تعبئة فتحات، يوجد فيها الفاعلون في العلاقات،

1- http://nlp.stanford.edu/software/relationExtractor.html

ويكون هدفها استخراج قيم المفعولين بهم في تلك العلاقات. قد يتسم بعض الفاعلين بالغموض بسبب كونهم يشيرون إلى عدة كيانات مختلفة موجودة في العالم الحقيقي. قد يحدث هذا الغموض بين الفئات المختلفة (قد يكون الجاغوار حيوانًا أو إحدى ماركات السيارات)، أو داخل الفئات (هناك الكثير من الأشخاص الذين يحملون اسم جون سميث). في الحالة الأخيرة على وجه الخصوص، من المفيد للغاية أن تكون معرفات الموارد الموحدة (URIs) موجودة كمُدخلات لكل فاعل من الفاعلين. على سبيل المثال، إذا كانت المهمة تتمثل في استخراج تواريخ الميلاد، تصبح النتيجة المتوقعة من عملية استخراج العلاقة نتيجة واحدة فقط لكل كيان فاعل، لكن عملية استخراج العلاقة ستعثر على الأرجح على أكثر من نتيجة واحدة لجون سميث. في حال وجود عدة معرفات موارد موحدة (URIs) مرتبطة بالاسم جون سميث في قاعدة المعرفة، فقد تستفيد عملية استخراج العلاقة من هذه المعلومات وتقوم بعرض عدة نتائج، وقد تحاول عرض تاريخ الميلاد الأكثر ترجيحًا لجون سميث المراد البحث عنه، وذلك في حال وجود معلومات أخرى عن أشخاص يحملون اسم جون سميث في قاعدة المعرفة، بناءً على تلك المعلومات الإضافية.

هناك عدد من قواعد البيانات متعددة المجالات، علمًا أن قاعدة بيانات DBpedia تمتلك أكبر عدد من الروابط التي تربطها بقواعد بيانات أخرى، وهو ما يجعلها من الناحية الفعلية بمنزلة مركز أو محور البيانات المترابطة. تشمل الأمثلة البارزة الأخرى لقواعد البيانات متعددة المجالات Freebase [82] وYago [83] وWikidata [84]. توجد قواعد بيانات محددة المجالات، وهي خاصة بعددٍ من المجالات المختلفة، فالحكومات تُصدر بياناتها باستخدام معايير الويب الدلالي، بينما تستفيد العلوم من الأساليب التكنولوجية لشرح العمليات المعقدة بواسطة الأنطولوجيات، فيما تقوم المكتبات والمتاحف بهيكلة وإصدار بياناتها الخاصة بالكتب والقطع الأثرية والوسائط، بينما يُثري مقدمو محتوى شبكات التواصل الاجتماعي مواقعهم بالمعلومات الدلالية. تعتمد إحدى طرائق استخراج العلاقات وهي طريقة الإشراف عن بعد (انظر القسم ٤-١٠)، على المخططات والبيانات المدرجة في قواعد البيانات المترابطة إلى حد بعيد.

من المهم معرفة أن المعلومات الموجودة في قواعد بيانات مختلفة غالبًا ما تكون مترابطة في مهمة استخراج العلاقات. قد يُعثر على معلومات تتعلق بالكيانات نفسها في أكثر من قاعدة بيانات واحدة، وللإشارة إلى ذلك، توجد في قواعد البيانات روابط تصل بينها. هذا يعني أن منهجيات استخراج العلاقات التي تستخدم المعلومات الموجودة أصلاً في قواعد البيانات قادرة على جمع المعلومات من قواعد بيانات عدة، كما سيتضح في وقت لاحق. علاوة على ذلك، هناك أيضًا روابط على مستوى المخططات (مثال: قد تكون الخاصية- تاريخ الميلاد» الموجودة في مخطط معين مرتبطة بالخاصية «مولود» في مخطط آخر، وقد تكون الفئة «ألبوم» مرتبطة بالفئة «ألبوم موسيقي»)، وهو ما يتيح سهولة أكبر في الجمع بين المعلومات الموجودة في قواعد البيانات، وأيضًا بين مخططات الاستخراج. على سبيل المثال، قد يُعرّف أحد المخططات أن الفنانين الموسيقيين لديهم تواريخ ميلاد، وقد يُعرّف مخطط آخر أنهم يقومون بإصدار الألبومات. يمكن إذًا الجمع بين هذه التعريفات لغرض استخراج كلتا العلاقتين.

٤-٦ أساليب استخراج العلاقات

بعد أن عرضنا طريقة عمل منهجية استخراج العلاقات النموذجية، سوف يشرح هذا القسم بالتفصيل مسارات استخراج العلاقات التي تعد بمنزلة أشكالٍ مختلفة لمنهجية استخراج العلاقات النموذجية التي ورد شرحها في القسم السابق. يمكن تقسيم منهجيات استخراج العلاقات بصفة عامة إلى أساليب قواعدية وأساليب خاضعة للإشراف وأساليب الاستخراج التمهيدي شبه الخاضعة للإشراف، وأساليب استخراج المعلومات غير الخاضعة للإشراف/ المفتوحة، والأساليب الخاضعة للإشراف عن بعد، والمخططات الشاملة.

٤-٦-١ منهجيات الاستخراج التمهيدي

كانت منهجيات الاستخراج التمهيدي، التي تعد نوعًا من المنهجيات شبه الخاضعة للإشراف، من أوائل منهجيات استخراج العلاقات، ومن أبرز الأساليب الرائدة في هذا الصدد طريقة استخراج علاقات الأنماط التكراري المزدوج (DIPRE) [85] ونظام Snowball [86]. فيما يلي وصف لطريقة DIPRE، لأن المنهجيات التي جاءت لاحقًا استخدمت بنيات هيكلية مماثلة.

تتكون منهجية طريقة DIPRE من أربع خطوات بسيطة (انظر إلى الخوارزمية 4-1). تشمل مُدخلات طريقة DIPRE المُدخل R، وهو عبارة عن مجموعة مكونة من خمس متواليات < s; o > للعلاقة PERSON author-of BOOK (شخص مؤلف كتاب)، والمُدخل D، وهو مجموعة وثائق، وفي هذه الحالة هذه المجموعة هي شبكة الإنترنت. تتمثل الخطوة الأولى في العثور على متواليات العلاقات الموجودة في شبكة الإنترنت. بعد ذلك تجري عملية توليد الأنماط. ثالثًا، يتم توليد الأنماط المطابقة. MD(P) هو مجموع متواليات العلاقات التي تكون أيٌّ من الأنماط p-P الموجودة فيها مطابقة للأنماط الموجودة في إحدى صفحات الإنترنت. تتكرر هذه العملية حتى يجري العثور على علاقات بعدد ن.

الخوارزمية ٤-١ DIPRE [85]: extract(R, D)
while R < n do (O ß findOccurrences(R, D (P ß generatePatterns(O (R ß MD(P end while return R

تُستخدم هذه الخوارزمية البسيطة تقريبًا في جميع منهجيات الاستخراج التمهيدي، مع اختلافات طفيفة. على سبيل المثال، قد يكون مُدخل الخوارزمية عبارة عن أمثلة وكذلك أنماط استخراج أو قواعد استخراج. يمكن إجراء عملية المطابقة بين الأنماط بطرق مختلفة، وذلك باستخدام عملية مطابقة دقيقة أو عملية مطابقة غير دقيقة. الجزء الأكثر إثارة للاهتمام في الخوارزمية هو طريقة توليد الأنماط. في منهجية DIPRE، تكون طريقة توليد الأنماط بسيطة للغاية، حيث يتم إنشاء نمط عن طريق تجميع الجمل التي تتطابق فيها سلسلة الكلمات بين كلمتي شخص وكتاب، والتي تظهر فيها الكلمتان شخص وكتاب بالترتيب نفسه. بعد ذلك، تقاس درجة الخصوصية، ففي حال مطابقة النمط لجُمل كثيرة؛ وكانت درجة الخصوصية فوق حد معين يُرمز له بالحرف t (تُضبط قيمته يدويًّا)، يُرفض النمط. أما إذا كانت درجة الخصوصية منخفضة جدًّا، ولم يُعثر إلا على الكتاب نفسه الذي يحتوي على ذلك النمط، يُرفض النمط أيضًا. هذا الأمر هو مؤشر يدل على أحد مساوئ منهجيات الاستخراج التمهيدي يعرف باسم المغزى

الدلالي، ويعني ذلك أن هذه المنهجيات تميل نحو الابتعاد كثيرًا عن المُدخل R وإنشاء أنماط تعبر عن علاقات مختلفة ذات صلة بعضها ببعض، وهي علاقات توجد غالبًا بصورة متوازية بجانب متواليات الكيانات ذاتها، على سبيل المثال، قد تتحول العلاقة من مؤلف كتاب إلى محرر كتاب.

جرى البحث في نماذج الاستخراج التمهيدي في وقت لاحق بهدف تحسين نموذج DIPRE. تشمل نماذج الاستخراج التمهيدي البارزة واسعة النطاق نماذج من قبيل نموذج KnowItAll [87] ونموذج NELL [88].

KnowItAll [87] هو نظام لاستخراج المعلومات يعتمد على سعة نطاق شبكة الإنترنت وتكرار معلوماتها لتوفير معلومات كافية والتحقق من صحتها. ونعني بالتكرار هنا أن كثيرًا من المعلومات المتاحة على الإنترنت توجد في أماكن متعددة في شبكة الإنترنت، وهو ما يعني أنه يمكن استخدام مصادر المعلومات المتعددة هذه من أجل التحقق من صحة الحقائق أو ملء الفجوات الناجمة عن المعلومات المفقودة. وبعكس نظام DIPRE، لا يبدأ نظام KnowItAll عمله انطلاقًا من علاقة واحدة، بل يبدأ بعدة علاقات، كما يحتوي على أساليب لتوسيع نطاق مخطط استخراج العلاقات. يتكون KnowItAll من أربع وحدات هي وحدة الاستخراج ووحدة واجهة محرك البحث ووحدة التقييم ووحدة الاستخراج التمهيدي.

تستخدم وحدة الاستخراج أنماط هيرست [89] من أجل استخراج النماذج الفردية لفئات الكيانات (هذه النماذج تكون نماذج فردية للفئة كتاب في نظام DIPRE). أنماط هيرست، التي سيتم شرحها في الفصل السادس، هي قواعد معجمية نحوية لاستخراج العلاقات، مثل NP1 هو NP2، حيث يشير NP2 إلى اسم فئة من فئات الكيانات مثل كتب، بينما يعني NP1 اسم النموذج الفردي لتلك الفئة. باستخدام واجهة محرك البحث، تُصاغ هذه الأنماط بعد ذلك (مع إبقاء NP1 فارغًا) على شكل استعلامات بحث من أجل استرجاع صفحات ويب تتضمن NP1. إضافة إلى ذلك، تضم هذه الوحدة قواعد لاستخراج العلاقات، على سبيل المثال، NP1 يلعب دورًا لصالح NP2، حيث تمثل هذه القاعدة العلاقة يلعب دورا لصالح (رياضي، فريق رياضي). بعد تطبيق جميع قواعد استخراج العلاقات، يجري التحقق من صحة الأنماط المستخلصة بواسطة وحدة التقييم.

تقوم وحدة التقييم بقياس إحصائيات التوارد المشترك للعلاقات التي يُحتمل استخراجها بواسطة عبارات مميزة، وتكون هذه العبارات المميزة على شكل أنماط استخراج عالية التكرار. هذا يعني أنه لكل استعلام من استعلامات البحث (مثال: توم كروز شارك في بطولة س)، يجري تدوين عدد نتائج البحث وحساب قيمة المعلومات المتبادلة الممثّلة بالنقاط (Pointwise Mutual Information [PMI]) للكيان توم كروز.

بعد ذلك يستخدم نظام KnowItAll عملية الاستخراج التمهيدي إلى جانب وحدة التقييم من أجل التحقق من صحة الأنماط المستخلصة. يجري استرجاع أعلى 20 نموذج فردي من حيث قيمة PMI وذلك لكل فئة من الفئات. بعد ذلك تُستخدم تلك النماذج الفردية لتدريب الاحتمالات الشرطية الخاصة بكل نمط من أنماط الاستخراج. تؤخذ بذور النماذج الفردية السالبة من النماذج الفردية الموجبة للفئات الأخرى. بعد ذلك يجري حفظ أفضل خمسة أنماط مستخلصة، فيما يجري التخلص من البقية. ثم يجري تدريب مُصنّف Naive Bayes الذي يجمع بين الأدلة المستقاة من تلك الأنماط الخمسة المستخلصة من أجل تصنيف ما إذا كان كيان معين (مثال: توم كروز) هو نموذج فردي لفئة معينة (مثال: ممثل). بدلاً من مجرد اختيار أفضل الأنماط المستخلصة مرة واحدة، يمكن استخدام عملية الاستخراج التمهيدي، أي أنه بمجرد تحديد أفضل خمسة أنماط مستخلصة، يمكن استخدامها للعثور على مجموعة جديدة من النماذج الفردية ذات قيمة PMI عالية. لضمان أن تكون جودة الأنماط المستخلصة مرتفعة، تُزال النماذج الفردية غير الصحيحة يدويًّا.

نظام NELL [88] هو نظام استخراج تمهيدي يستخلص المعلومات من شبكة الإنترنت من أجل تعبئة قاعدة معرفة، وبمرور الوقت، يتعلم كيفية استخراج المعلومات بدقة أعلى. وكما هو الحال مع نظام KnowItAll، يعدُّ نظام NELL مبنيًّا على فرضية أن المعلومات الضخمة عالية التكرار الموجودة على شبكة الإنترنت هي بمنزلة ميزة هائلة يمكن لآليات التعلم الاستفادة منها. تكمن الاختلافات الرئيسة بين النظامين في أن وحدة الاستخراج التمهيدي هي أكثر تعقيدًا في الأخير، وأن نظام NELL يجمع بين الأنماط المستخلصة من مصادر مختلفة على شبكة الإنترنت، بما فيها النصوص والقوائم والجداول. ومثل نظام KnowItAll، يتعلم هذا النظام كيفية استخراج أي النماذج

الفردية تنتمي لأي الفئات، وأي العلاقات توجد بين النماذج الفردية لتلك الفئات.

يتم استخراج المعلومات من معلومات غير مهيكلة موجودة على شبكة الإنترنت (نص)، ومن بيانات شبه مهيكلة (قوائم وجداول). تُدرّب أدوات استخراج المعلومات بصورة متناسقة باستخدام التعلم المقترن، وذلك باستخدام نظام CPL للنص الحر ونظام CSEAL للقوائم والجداول [90]. ومثل نظام KnowItAll، يعتمد نظام CPL على إحصائيات التوارد المشترك بين أشباه الجمل الاسمية وأنماط النص من أجل تعلم أنماط الاستخراج. يستخدم نظام CSEAL علاقات الاستبعاد المتبادل لتوفير أمثلة سلبية، وهو ما يُستخدم بعد ذلك لفلترة القوائم والجداول التي تتسم بالعمومية المفرطة.

إضافة إلى ذلك، يتعلم نظام NELL الانتظام الصرفي للنماذج الفردية لفئات الكيانات، ويستخدم قواعد عبارات هورن الاحتمالية بغية استنتاج علاقات جديدة من العلاقات التي سبق له تعلمها. ولتعلم الانتظام الصرفي، يستخدم نظام NELL مُصنّفًا صرفيًّا مقترنًا (CMC). لكل فئة من الفئات، يجري تدريب نموذج لوجستي تراجعي لتصنيف العبارات الاسمية بناءً على خصائصها الصرفية والنحوية (مثال: نوع الكلمات واستخدام الأحرف الكبيرة والسوابق واللواحق وبطاقات تصنيف أقسام الكلام). يتدرب مُتعلم القواعد عبارات هورن من أجل استنتاج علاقات جديدة من العلاقات الموجودة أصلاً في قاعدة المعرفة.

يبدأ نظام التعلم بإحدى قواعد المعرفة (١٢٣ فئة، ٥٥ علاقة، وبضع نماذج فردية للفئات وثلاثيات العلاقات)، ومن ثمّ يبدأ بتعبئة قاعدة المعرفة وزيادة حجمها بصورة تدريجية. بعد قيام وحدة الاستخراج باستخراج اعتقاد ما، يبدأ تحسين دقة هذا الاعتقاد عبر الرجوع إلى مصادر بيانات خارجية أو أشخاص متخصصين. بعدها تُرفع الاعتقادات المدعومة بقوة أكثر من غيرها إلى مرتبة حقائق، وتُدمج في قاعدة المعرفة. في بقية خطوات الاستخراج، تستخدم وحدة الاستخراج دائمًا قاعدة المعرفة التي جرى تحديثها.

يوفر نظام NELL في العادة إمكانية استخراج النماذج الفردية للفئات وكذلك العلاقات بدقة عالية نسبيًّا في بداية الأمر [88]، وعادة ما تكون مكونات الاستخراج المختلفة مكملا بعضها بعضًا. ومع ذلك فهي تشير إلى مشكلة تعد شائعة في منهجيات

الاستخراج التمهيدي، وهي ضعف دقة الاستخراج مع مرور الوقت. غير أنه من الممكن حل هذه المشكلة عبر السماح للعنصر البشري بالتفاعل مع النظام أثناء عملية التعلم، وذلك باستخدام أسلوب التعلم النشط [91].

٤-٧ المنهجيات المعتمدة على القواعد

هناك أسلوب آخر لإنشاء أنظمة استخراج العلاقات، وهو استخدام منهجية قواعدية أو نمطية. تستفيد المنهجيات القواعدية لاستخراج العلاقات من المعرفة المجالية (أو المعرفة بالمجال)، ويجري ترميز هذه المعرفة المجالية على شكل قواعد لاستخراج العلاقات [92–94]. هناك نوعان مختلفان من المنهجيات القواعدية، وهما المنهجيات المنفصلة والمنهجيات التي تتعلم القواعد لغرض الاستدلال بهدف تكملة منهجيات استخراج العلاقات الأخرى. يعتمد النوع الأول عادة على قواعد نحوية لترميز القواعد المعقدة وعلاقات التبعية الموجودة بينهما. من الأمثلة على الأشكال القواعدية عضو فرقة موسيقية تليه بعد 30 حرفًا أو أقل آلة موسيقية. للتعرف على عضو الفرقة الموسيقية والآلة كليهما، تُستخدم معاجم كيانات أسماء مسبقة التجميع وكذلك التعبيرات العادية. من مساوئ مثل هذه المنهجيات القواعدية كونها غير قادرة على تعميم قدرتها على التعرف لتشمل الأنماط النصية غير المرئية، إلى جانب ضعف قدرتها على الاستدعاء.

تتضمن المنهجيات القواعدية المستخدمة لأغراض الاستدلال نظام KnowledgeVault [95]، الذي يستخدم خوارزمية ترتيب تعتمد على المسارات. تبدأ العملية بزوجين من الكيانات يُعرف أن بينهما علاقة وفقًا لقاعدة معرفة (بذرة)، وبعد ذلك يسير النظام بطريقة عشوائية فوق خط المعرفة لإيجاد مسارات أخرى تربط بين هذه الكيانات. لذا يمكن أن يتعلم النظام هل يوجد طفلٌ مشتركٌ بين شخصين أم لا، أو هل هناك احتمال كبير في أن يتزوج هذان الشخصان أم لا، أو أن الأشخاص غالبًا ما يدرسون في الجامعة نفسها التي يدرس فيها أشقاؤهم. من مساوئ استخدام القواعد التي جرى تعلمها لأغراض الاستدلال أن القواعد التي جرى تعلمها بواسطة قاعدة معرفة صغيرة قد لا تكون عامة بما يكفي لتنطبق على علاقات جديدة، على سبيل المثال، يتسبب استخدام مثل هذه القواعد المكتسبة عن طريق التعلم في حدوث انخفاض في

الأداء [75] في بعض التجارب التي قدمت في مؤتمر تحليل النصوص - تعبئة قواعد المعرفة (TAC KBP) في عام ٢٠١٤م. وللتخفيف من هذه المشكلة، ينبغي الحرص على استخدام القواعد التي تعتمد على إثباتات كافية.

٤-٨ المنهجيات الخاضعة للإشراف

تعد المنهجيات الخاضعة للإشراف في الوقت الراهن أفضل منهجيات استخراج العلاقات من حيث الأداء، شريطة وجود ما يكفي من البيانات التدريبية المصنفة. تسير هذه المنهجيات بدقة وفقًا للمنظومة العامة لاستخراج العلاقات (الشكل 1-4)، حيث تقوم باستخدام مكنز أضيفت له الحواشي والتعليقات لإجراء عملية المعالجة المسبقة للجُمل بواسطة خطوات المعالجة المسبقة المعتادة في عمليات معالجة اللغات الطبيعية (تصنيف أقسام الكلام، التحليل الإعرابي، تحديد كيانات الأسماء، ...الخ)، وبعد ذلك تقوم باستخراج الخصائص وتدريب أحد النماذج والتنبؤ بالعلاقات في مجموعة من بيانات الاختبار.

تُستخلص الخصائص من الأمثلة الإيجابية والسلبية على حدٍّ سواء، وتكون الخصائص بمنزلة إشارات تتيح تعلم ما إذا كانت هناك علاقة ما بين كيانين من كيانات الأسماء أو لا. أثناء عملية التدريب، يلاحظ النموذج مدى تكرار ورود خاصية معينة باستخدام أمثلة إيجابية مقابل أمثلة سلبية، وبناءً على ذلك يتعلم وزن كل خاصية من الخصائص، وهذا الوزن يمكن أن يكون إيجابيًّا أو سلبيًّا. على سبيل المثال، إذا كانت العبارة الفاصلة بين كيانين تجمعهما العلاقة مؤلف [كتاب] هي عبارة هو مؤلف [كتاب]، فسوف تُعطى وزنًا إيجابيًّا مرتفعًا، بينما تحصل العبارة هو مدير على وزن سلبي.

تشمل الخصائص المعتادة في عملية استخراج العلاقات (المستخدمة على سبيل المثال في [73 ,81]) الآتي:

- N-gram من الكلمات الموجودة على يسار ويمين الكيانات؛
- N-gram من أقسام الكلام التي تنتمي إليها الكلمات الموجودة على يسار ويمين الكيانات؛
- علامة تشير إلى أول كيان يرد في الجملة؛

- سلسلة بطاقات تصنيف أقسام الكلام وكيس الكلمات (BOW) بين الكيانين؛
- مسار التبعية بين الفاعل والمفعول به؛

بطاقات تصنيف أقسام الكلام التي تنتمي إليها الكلمات الموجودة في مسار التبعية بين الكيانين؛ والجذوع الموجودة في مسار التبعية.

تشمل الخصائص الأخرى الممكنة أساليب كيرنيل [96 ,97] أو تضمينات العلاقات التي ظهرت مؤخرًا، والتي تتعلم تمثيلات ذات أبعاد أعلى للبيانات المصنّفة. يمكن اعتبار هذه التمثيلات كخصائص كامنة ولذا تزول الحاجة لعملية هندسة الخصائص التي قد تكون مرهقة. من حيث النماذج، يجري استخدام تشكيلة واسعة مثل SVM أو نماذج الإنتروبيا القصوى أو شبكات ماركوف المنطقية أو الشبكات العصبية (العميقة).

من الأمثلة على أدوات استخراج العلاقات النموذجية أداة Stanford لاستخراج العلاقات[1]، المبنية كوحدة إضافية على منصة Stanford CoreNLP. تكتشف هذه الأداة بعض العلاقات من قبيل (يعيش في، يوجد في، يوجد مقر المؤسسة في، ويعمل في). هذه الأداة مدربة بواسطة بيانات مكنز TREC، لكن من السهل إعادة تدريبها باستخدام مكنز آخر وتخصيصها.

٤-٩ المنهجيات غير الخاضعة للإشراف

باتت المنهجيات غير الخاضعة للإشراف لاستخراج العلاقات تحظى بالشعبية بعد فترة وجيزة من ظهور الأنظمة الخاضعة للإشراف، وكان من بين الأمثلة على أنظمة استخراج المعلومات المفتوحة أنظمة من قبيل TextRunner [99] وReVerb [100] وOLLIE [101]. يستخدم منهج أنظمة استخراج المعلومات المفتوحة أساليب بسيطة وقابلة للتوسيع لاستخراج المعلومات غير المقيّدة مسبقًا. هذا المنهج هو عكس المنهجيات شبه الخاضعة للإشراف التي سبق شرحها في الأقسام السابقة، والتي تستخدم مخططات استخراج معرّفة مسبقًا. لذا يمكن اعتبار أنظمة استخراج المعلومات المفتوحة كمجموعة فرعية من المنهجيات غير الخاضعة للإشراف. هذا يعني أنه يتعين على أنظمة استخراج المعلومات المفتوحة استنتاج الفئات التي تنتمي إليها

1- http://ai.cs.washington.edu/projects/open-information-extraction

الكيانات، وكذلك العلاقات القائمة بينها. فيما يلي شرح لأول منهجية من منهجيات استخراج المعلومات المفتوحة، وذلك من أجل استعراض مسارات الأبحاث والإشارة إلى أوجه القصور والتحسينات الموجودة في الأبحاث اللاحقة.

كان نظام TextRunner [99] أول نظام مفتوح لاستخراج المعلومات يجري تطبيقه وتقييمه بالكامل. يتعلم هذا النظام نموذج حقل شرطي عشوائي (CRF) للعلاقات وفئات الكيانات والكيانات، ويتعلم هذا النموذج من أحد المكانز بواسطة نموذج استخراج لا يعتمد على العلاقات. أولاً، يقوم النظام بمعاينة المكنز بأكمله مرة واحدة، ويقوم بإضافة التعليقات والحواشي إلى الجمل ببطاقات تصنيف أقسام الكلام وأشباه الجمل الاسمية. لتحديد ما إذا كان ينبغي استخراج العلاقة أم لا، يستخدم النظام أداة تصنيف خاضعة للإشراف. هذه الأداة مدربة عن طريق إجراء تحليل إعرابي لمجموعة فرعية صغيرة من المكنز، ومن ثمّ تصنيف الجمل وفق منهج تجريبي إلى أمثلة إيجابية (موثوقة) وسلبية (غير موثوقة)، وذلك باستخدام مجموعة محدودة من القواعد المشفرة يدويًّا. بعد ذلك تقوم أداة التصنيف باتخاذ قرار بشأن الجمل غير المرئية بناءً على بطاقات تصنيف أقسام الكلام بدلاً من شجرة الإعراب، لأن عملية التحليل الإعرابي للمكنز بأكمله باهظة الثمن. للتمييز بين المترادفات، يقوم نظام TextRunner بإجراء عملية تجميع غير خاضع للإشراف للعلاقات والكيانات بناءً على أوجه الشبه من حيث التسلسل والتوزيع [99].

يعالج نظام ReVerb [100] اثنين من أوجه القصور الموجودة في أنظمة استخراج المعلومات المفتوحة القديمة، وهما عدم تناسق المعلومات المستخلصة وعدم احتوائها على معلومات مفيدة. تحدث مشكلة عدم تناسق المعلومات المستخلصة عندما تفتقر شبه الجملة الاسمية المستخلصة إلى تفسير ذي معنى. يعود السبب إلى حقيقة مفادها أن القرارات تتخذ بشكل تسلسلي في نظام TextRunner. من الأمثلة على ذلك العلاقة (يحتوي، يُغفل) التي تُستخلص من الجملة (الدليل يحتوي على روابط لا تعمل ويُغفل المواقع الإلكترونية). لحل هذه المشكلة، تفرض قيود نحوية على العلاقات التي ينبغي استخراجها. أول هذه القيود أنه ينبغي أن تكون شبه جملة العلاقة إما بصيغة الفعل (مثال: اخترع) أو بصيغة فعل متبوع بحرف جر (مثال: يوجد في) أو بصيغة فعل متبوع بأسماء أو صفات أو ضمائر وحرف جر (مثال: يصل وزنه الذري إلى). أيضًا، إذا كانت

هناك عدة تطابقات ممكنة، يجري اختيار التطابق الأطول. في حال العثور على تسلسلات متجاورة (مثال: يرغب، في تمديد)، تُدمج هذه التسلسلات (مثال: يرغب في تمديد). أخيرًا، يجب أن تظهر العلاقة بين المعطيين في الجملة.

تُغفِل المعلومات المستخلصة غير المفيدة معلومات مهمة، على سبيل المثال، يستخلص نظام TextRunner فوست، عقد، صفقة بدلاً من استخراج فوست، عقد صفقة مع، الشيطان، من الجملة فوست عقد صفقة مع الشيطان. يمكن استخراج بعض المعلومات المفقودة بواسطة القيود النحوية. غير أن ذلك قد يسبب استخراج علاقات مفرطة في درجة التحديد، على سبيل المثال: لا يقدم سوى أهداف متواضعة لخفض غازات الاحتباس الحراري في. لحل هذه المشكلة، يُستحدث قيد معجمي يتمثل في ضرورة أن تظهر العلاقة مع 20 معطى من المعطيات المتمايزة على الأقل في الجملة لكي تكون مفيدة.

وعلى الرغم من كون مجال استخراج المعلومات من مجالات البحث الواعدة، وعلى الرغم من إمكانية رسم خريطة لمجموعات العلاقات تتوافق مع مخططات استخراج العلاقات لاحقًا، إلا أن ذلك يضع قيودًا غير ضرورية على مهمة تعبئة قواعد المعرفة. يمكن توقع أن يكون أداء منهجيات استخراج العلاقات المطورة لمخطط معين أعلى من أداء منهجية غير محصورة بمخطط معين. والسبب في ذلك يعود إلى المشكلات المذكورة أعلاه والمتمثلة في العلاقات غير المتناسقة وغير المفيدة. تكون حدة هذه المشكلات أقل في أساليب الاستخراج التمهيدي.

على الجانب الآخر، تعد أساليب استخراج المعلومات المفتوحة التي لا تستخدم مخططات معرّفة مسبقًا قابلة للتطبيق بشكل أوسع في سيناريوهات مختلفة. من الأمور التي يمكن اعتبارها كمزايا إمكانية تحويل المخرجات، حسب السيناريو، إلى مخططات مختلفة في خطوة تأتي في مرحلة ما بعد المعالجة. تتوفر الأساليب والنماذج التجريبية لأنظمة استخراج المعلومات المفتوحة بصورة منفصلة عن مشروع KnowItAll من جامعة واشنطن (TextRunner، ReVerb، Ollie، Srlie، Relnoun)،[1] وتم نشرها من قبل باحثي Stanford NLP، وهي مدمجة بمنصة Stanford CoreNLP [102][2].

1- http://nlp.stanford.edu/software/openie.html

2- http://nlp.stanford.edu/software/mimlre.shtml

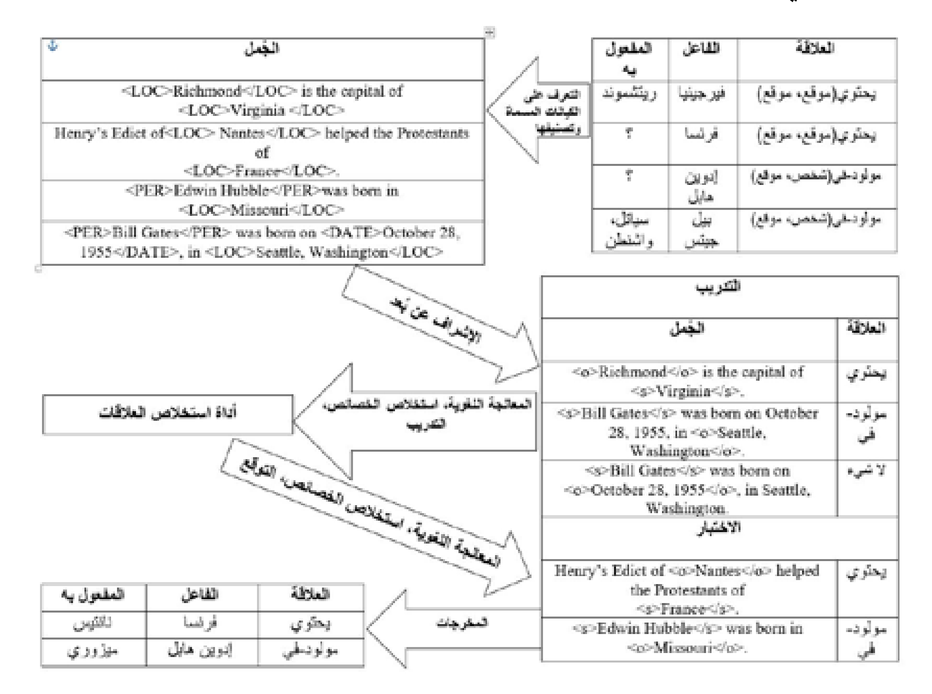
٤-١٠ منهجيات الإشراف عن بُعد

الإشراف عن بعد هو أسلوب لإضافة التعليقات والحواشي للبيانات التدريبية بصورة آلية باستخدام العلاقات الموجودة في قواعد المعرفة. في عام 1999م، اقترح كريفين وكولين [103] المنهجية الأولى كأسلوب لتعبئة قواعد المعرفة في مجال الطب الحيوي، على الرغم من إطلاقهما تسمية «ضعيف التصنيف» على منهجيتهما. وعلى الرغم من كون النتائج واعدة، إلا أن هذه المنهجية لم تحظ بالشعبية إلا بعد مرور ١٠ سنوات، وذلك عندما استحدث مصطلح «الإشراف عن بعد». قد يعود سبب بروز هذه المنهجيات على السطح مرة أخرى جزئيًّا إلى زيادة توفر قواعد معرفة على شبكة الإنترنت. يُعرّف (مينتز وآخرون) [81] فرضية الإشراف عن بعد كالتالي:

> في حال مشاركة كيانين في علاقة ما، يمكن أن تعبر أي جملة تحتوي على هذين الكيانين عن تلك العلاقة.

يقدم الشكل ٤-٢ صورة لكيفية عمل مثل هذه المنهجية. يتمثل مُدخل هذه المنهجية في قاعدة معرفة تحتوي على مجموعة من فئات الكيانات والعلاقات، ونماذج لتلك الفئات وأمثلة على تلك العلاقات، وكذلك مكانز تدريب واختبار. تجري معالجة مكنز التدريب مسبقًا بغية التعرف على كيانات الأسماء، وبعدها يجري البحث فيه عن جُمل تحتوي على الفاعل والمفعول به كليهما في العلاقات المعروفة (مثال: فيرجينيا وريتشموند في العلاقة تتضمن (موقع، موقع)). تعدُّ الجمل التي تحتوي على الفاعل والمفعول به كليهما في العلاقات المعروفة بيانات تدريب إيجابية في العلاقة، بينما تعدُّ الجمل الأخرى أمثلة تدريب سلبية (NIL). بعدها يجري تدريب أداة تصنيف خاضعة للإشراف (مثال: Naive Bayes، SVM، MaxEnt) وتطبيقها على مكنز اختبار. بصورة عامة، تكون عملية التعلم مطابقة لعملية التعلم المستخدمة في أنظمة التعلم الخاضع للإشراف، وليس هناك اختلاف سوى في عملية تصنيف بيانات التدريب (تكون العملية آلية بدلاً من أن تكون يدوية). لذا، تحتوي هذه المنهجية على جميع مزايا التعليم الخاضع للإشراف (دقة عالية في المخرجات المستخلصة بالنسبة إلى مخطط الاستخراج)، ومزايا إضافية، لأنه ليس من المطلوب بذل مجهود يدوي في تصنيف بيانات التدريب. يكون أداء عملية الاستخراج أدنى قليلاً من أداء المنهجيات الخاضعة للإشراف، وذلك بسبب

تصنيف أمثلة التدريب بصورة خاطئة. من الأسباب الرئيسة المؤدية إلى تصنيف أمثلة التدريب بصورة خاطئة غموض الأشكال السطحية (مثال: فرجينيا يمكن أن تكون اسم شخص أو موقع) [105, 104]. ظلت مسألة تحسين عملية التصنيف الآلي لأمثلة التدريب في محور الاهتمام في بحوث منهجيات الإشراف عن بُعد منذ ذلك الوقت، كما هو مذكور في استبانة أجراها [106].



الشكل ٤-٢: [٨١] نظرة عامة على أسلوب الإشراف عن بُعد.

٤-١٠-١ المخططات الشاملة

يجمع مفهوم المخططات الشاملة [107] بين مزايا عمليتي استخراج المعلومات المفتوح والإشراف عن بُعد. تفترض طرق نمذجة البيانات المفقودة لتقليل النتائج الخاطئة أنه لا يتم تضمين جميع العلاقات (مثال: مايكروسوفت أسسها بيل جيتس)، وهو ما يؤدي إلى تصنيفها كبيانات تدريب سلبية. على الجانب الآخر، تتناول المخططات الشاملة مفهوم أن ليس جميع العلاقات (مثال: أسسها) موجودة في قاعدة المعرفة. بعد

ذلك تسعى إلى الجمع بين العلاقات المعرّفة بواسطة مخطط قاعدة المعرفة والعلاقات المكتشفة في النص باستخدام أساليب استخراج المعلومات المفتوح. نشير هنا أن أساليب استخراج المعلومات المفتوح لا يعتمد على مخطط استخراج، بل يقوم بتجميع الأنماط السطحية (مثال: أسسها، قام بتأسيسها) بدلاً من ذلك على شكل علاقات. ولإجراء ذلك، يتم تكوين مصفوفة تمثل صفوفها أزواج الكيانات وتمثل أعمدتها كلتا العلاقتين المعرفتين في قاعدة المعرفة وأنماط استخراج المعلومات المفتوح. ولتوقع قيم العلاقات غير المرئية، يتم استخدام طريقة تعميل (أي التحليل إلى عوامل) المصفوفة.

٤-١٠-٢ المنهجيات الهجينة

أخيرًا، تجدر الإشارة إلى أنه بالإضافة إلى المخططات الشاملة، هناك عدد كبير من المنهجيات الهجينة الموجهة نحو الجمع بين مزايا عدة أنواع من المنهجيات. هناك أساليب تجمع بين المنهجيات الهجينة القائمة على الأنماط والمنهجيات الخاضعة للإشراف، والمنهجيات التي تجمع بين منهجيات الإشراف عن بُعد والمنهجيات القواعدية [108]، والمنهجيات التي تجمع بين الإشراف عن بُعد والإشراف (المباشر) [109]، وأخيرًا، الأساليب التي تجمع بين المخططات الشاملة والمنهجيات القواعدية [110].

من أدوات استخراج العلاقات الجديدة التي تحظى بالشعبية أداة SampleJS [109][1]. تستخدم هذه الأداة الإشراف عن بُعد للحصول على أمثلة تدريبية مشوشة، وتستخدم التعليم النشط لتحسين جودة البيانات التدريبية بصورة تكرارية. تعالج هذه المنهجية بعض المشكلات التي ورد شرحها في المقدمة، على سبيل المثال العلاقات التي يمكن أن تتداخل. يأتي هذا التوزيع مرفقًا بنموذج مسبق التدريب يستخدم مزيجًا من مخطط العلاقات Freebase ومخطط TAC KBP 2013، وهو ما ينتج عنه ٤١ علاقة، كما تُستخدم موسوعة ويكيبيديا كمكنز تدريبي. يمكن إعادة تدريب هذه المنهجية للمخططات و/ أو المكانز الأخرى.

1- http://www.nzdl.org/vikification/docs.html

٤-١١ الأداء

هناك عدة مكانز تدريبية لعملية استخراج العلاقات الخاضعة للإشراف، على الرغم من أن عددها لا يقترب من عدد المكانز المتوفرة لعملية التعرف على كيانات الأسماء. تشمل المكانز ACE وOntonotes وTREC وTAC KBP. تشمل مكانز ACE و Ontonotes أيضًا تعليقات وحواشيَ لمهام معالجة اللغات الطبيعية المترابطة، مثل مهمة التعرف على كيانات الأسماء واستخراج الإحالات المشتركة، وهو ما يجعلها مثالية لدراسة الاعتماد المتبادل بين تلك المهام.

يعتمد أداء منهجيات استخراج العلاقات اعتمادًا كبيرًا على نوع العلاقة. عندما يتعلق الأمر بالمنهجيات المبنية على التعلم، يعتمد الأداء على عدد الأمثلة التدريبية الموجود لكل علاقة، وبالنسبة للمنهجيات التي تستخدم المعرفة الأساسية مثل منهجيات الإشراف عن بُعد والمنهجيات القواعدية، يعتمد الأداء على جودة البيانات الأساسية وكذلك على نوع نص المكنز (مثال: النصوص الإخبارية، نصوص ويكيبيديا، بيانات الطب الحيوي). تتيح مبادرات التقييم من قبيل مؤتمرات تحليل النصوص - تعبئة قواعد المعرفة TAC KBP لتقييم أساليب تعبئة إجراء مقارنة موضوعية بين المنهجيات المختلفة عبر استخدام بعض من هذه العوامل كمتغيرات تحكم. في مؤتمر TAC KBP لعام ٢٠١٤، استخدمت المقترحات المقدمة جميع أنواع منهجيات استخراج العلاقات المختلفة التي نوقشت في هذا الفصل، ونعني بذلك منهجية الإشراف المباشر ومنهجية الإشراف عن بُعد والمنهجيات المبنية على الأنماط والمنهجيات المبنية على القواعد، ومنهجيات الاستخراج التمهيدي ومنهجيات استخراج المعلومات المفتوح ومنهجيات المخططات الشاملة. تشير الاتجاهات الناشئة إلى أن ١٤ من أصل ١٨ نظامًا قُدمت إلى المؤتمر استخدمت منهجيات الإشراف عن بُعد، وأن معظم الأنظمة جمعت بين الإشراف عن بُعد والقواعد، بالإضافة إلى أن أهم ثلاثة أنظمة كانت مبنية على منهجية الإشراف عن بُعد. يعد التعلم النشط أسلوبًا ناجحًا للجمع بين منهجيتي الإشراف المباشر والإشراف عن بُعد، علمًا أن إحدى هاتين المنهجيتين تشكل أساس أداة SampleJS [109]. قدمت المنهجية الوحيدة المستندة إلى المخططات الشاملة أداءً جيدًا، على الرغم من أن أداءها لم يكن بدرجة أداء منهجية الإشراف عن بُعد المدمجة نفسها، إما مع منهجية الإشراف

المباشر أو المنهجية القواعدية. كان أداء المجموعات التي استخدمت أنماطًا مصنوعة يدويًّا إما متوسطًا أو دون المتوسط، وكان أفضل منهجية من بين تلك المنهجيات تلك التي جمعت بين استخراج المعلومات المفتوح والأنماط المصنوعة يدويًّا. يشير هذا الأمر إلى أنه عندما يتعلق الأمر بتعبئة قواعد المعرفة بالعلاقات، فإن المنهجيات المستندة إلى التعلم الآلي تتفوق بشكل كبير على المنهجيات المستندة إلى الأنماط. بصفة عامة، وصل الأداء البشري في مؤتمر TAC KBP لعام ٢٠١٤ إلى درجة F1 نسبتها ٣٦,٧٠٪، في حين حققت المنهجية الأفضل من حيث الأداء نسبة ٣٦,٧٧٪.

من مساوئ مؤتمر تحليل النصوص - تعبئة قواعد المعرفة (TAC KBP) أن عدد أمثلة التدريب لكل علاقة يختلف اختلافًا واسعًا، وهو ما يجعل من الصعب إجراء مقارنة بين أداء العلاقات. لإعطاء لمحة عن صعوبة عملية استخراج العلاقات، يضم الجدول ٤-١ قائمة درجات P وR وF1 الخاصة العلاقات الأكثر شيوعًا في نظام التقييم SampleJS [109]، وهي مكونة جزئيًّا من علاقات مؤتمر TAC KBP لعام 2014م وجزئيًّا من علاقات قاعدة المعرفة Freebase.

الجدول ٤-١ : مقارنة بين أداء العلاقات المختلفة

الأسلوب	P	R	F1
موظف في	32	46	38
أهم الأعضاء	26	60	36
(Org:) alt names	48	39	43
اللقب	26	35	30
الزوج(ة)	54	85	66
الأصل	43	70	53
سبب الوفاة	93	39	55
الأطفال	62	18	27
تاريخ الوفاة	64	39	48
السن	97	90	93

كما يظهر من الجدول، يختلف الأداء اختلافًا واسعًا حسب نوع العلاقة، على سبيل المثال، يكون أداء F1 في علاقة السن ٩٣٪، في حين لا يكون هذا الأداء في علاقة الأطفال سوى ٢٧٪. تجدر الإشارة إلى أن تحديات التقييم هذه لا تعطي بالضرورة فكرة واقعية عن أداء عمليات استخراج العلاقات في التطبيق العملي. يزيد أداء عملية استخراج العلاقات بصورة درامية مع وجود بيانات تدريب إضافية، وأيضًا عند التخلص من العلاقات التي جرى اسخلاصها بمستوى ثقة منخفض. نجح نظام هجين على مقياس شبكة الإنترنت لاستخراج العلاقات أنشأته شركة جوجل [95] في تحقيق درجة AUC (منطقة تحت منحنى استدعاء-الدقة area under the precision-recall curve) قيمتها ٩٢٧,٠، وذلك عبر التخلص من جميع العلاقات المستخلصة بمستوى ثقة يقل عن ٩,٠.

باختصار، تتمثل منهجيات استخراج العلاقات الأكثر نجاحًا في المنهجيات الهجينة التي تجمع بين المنهجيات المستندة إلى التعلم التي تستخلص المعلومات باستخدام عدد من الأساليب المختلفة. تستخدم هذه المنهجيات كميات كبيرة من بيانات التدريب وتستخلص العلاقات من عدة مصادر مختلفة.

٤-١٢ خلاصة

يلخص الجدول ٤-٢ النقاط الرئيسة المتعلقة بأنواع المنهجيات المختلفة. توجد في جميع أساليب استخراج العلاقات مزايا وعيوب، فهي تختلف في كمية المدخلات الأولية المطلوبة، وما إذا كانت هناك حاجة للتدخل البشري أو لا أثناء عملية التعلم، ومدى ملاءمتها لعملية تعبئة قواعد المعرفة. قد لا تحتاج أساليب الاستخراج التمهيدي سوى بضعة أمثلة من الأمثلة الأولية، لكن كما نوقش في القسم ٤-٦-١، قد تتطلب مشكلة المغزى الدلالي مزيدًا من التدخل البشري أثناء عملية التعلم. تعد هذه الأساليب ملائمة لتعبئة قواعد المعرفة، نظرًا لأن عملية الاستخراج تجري وفقًا لمخطط استخراج. تتطلب المنهجيات القواعدية عددًا كبيرًا من القواعد المطورة يدويًّا، بالإضافة إلى معاجم جغرافية لكيانات الأسماء، وعادة ما تكون قدرة الاستدعاء لديها متدنية. في سيناريوهات التطبيق العملي، لا تزال المنهجيات القواعدية تستخدم في أحيانٍ كثيرة، على الرغم من أنها لا تعدُّ حديثة من ناحية الأداء. يعود السبب في ذلك إلى

سهولة تطويرها وتوسيعها، ولا تتطلب بذلك جهدًا كبيرًا مسبقًا، مثل تصنيف مكنز تدريب. هناك صيغة لمنهجية استخراج العلاقات القواعدية لا تتطلب بذل جهد، وهي الأنظمة المتعلمة للقواعد، والتي بدورها تتعلم قواعد استنتاج عالية الدقة باستخدام بذور قواعد المعرفة، والتي يمكن استخدامها إلى جانب أساليب استخراج العلاقات الأخرى.

تتطلب أساليب استخراج العلاقات الخاضعة للإشراف أمثلة تدريبية مصنفة وفقًا لمخطط علاقات. تعد هذه الأساليب أفضل أساليب استخراج العلاقات في تعبئة قواعد المعرفة، إلا أنها قد تتطلب أيضًا بذلك جهدًا كبيرًا مسبقًا في حال عدم توفر بيانات تدريبية مناسبة. لا تتطلب منهجيات استخراج المعلومات المفتوحة أي مُدخلات في البداية، لكن هذا يعني أن مخرجات مثل هذه المنهجيات لا تكون سوى تجميعات للعلاقات، وليس هناك طريقة بسيطة لتحويلها إلى مخطط علاقات موجود سابقًا. لذا، تعدُّ هذه المنهجيات محل اهتمام في السيناريوهات التي لا تتوفر فيها مخططات علاقات، أو التي يكون هدفها توسيع نطاق أحد مخططات العلاقات، لكنها أقل ملاءمة لعمليات تعبئة قواعد المعرفة.

تتطلب منهجيات الإشراف عن بُعد كمية صغيرة من المُدخلات، نحو 30 مثالاً لكل علاقة على الأقل، وتستخدم هذه المعلومات لتصنيف بيانات التدريب، ومن ثمّ إجراء عملية التعلم الخاضع للإشراف. بسبب وجود مثل هذه المعلومات بوفرة على شبكة الإنترنت ضمن قواعد بيانات موجودة حاليًّا، تصبح عملية جمع هذه المعلومات آليًّا أمرًا ممكنًا، ولذا فإنها لا تتطلب العامل البشري. ونظرًا لأنها أيضًا تستخدم بعد ذلك المخططات المرتبطة بأمثلة العلاقات الخاصة بالتدريب، فإنها تعد مناسبة للغاية لعمليات تعبئة قواعد المعرفة. وحتى لو توفرت معلومات تدريب مصنفة، كما هو الحال في حملات التقييم من قبيل مؤتمرات TAC KBP، فإن الأداء يتحسن عند إضافة بيانات إضافية مصنفة عن بُعد. تعدُّ المخططات الشاملة منهجية تقوم بتوحيد العلاقات المُعرّفة بواسطة المخططات. يمكن استخراج هذه العلاقات باستخدام أساليب مختلفة لاستخراج العلاقات، مثل الإشراف عن بُعد واستخراج المعلومات المفتوح، وهذا من نقاط القوة الرئيسة الموجودة فيها.

إذًا، تحديد الأسلوب الأمثل لاستخراج العلاقات يعتمد في حقيقة الأمر على المهمة المطروحة. إذا كانت المهمة استكشافية، يكون أسلوب استخراج المعلومات المفتوح ملائمًا بقوة، وهناك العديد من الأدوات التي تتيح معرفة أدائها. بالنسبة لعمليات تعبئة قواعد المعرفة، تتكون الوسائل الحديثة المستخدمة حاليًّا من منهجيات هجينة تجمع بين أساليب استخراج المعلومات الخاضعة للإشراف، وأساليب الإشراف عن بُعد أو القواعد المستنتجة باستخدام بذور قواعد المعرفة.

الجدول ٤-٢: مقارنة الحد الأدنى بين طرق استخلاص المعلومات الخاضعة للإشراف

المنهجية	المدخلات	المخرجات	الوصف	المزايا	العيوب
الاستخراج التمهيدي	نص غير مصنف و/ أو مخططات علاقات و/ أو قواعد و/ أو أمثلة	قواعد استخراج وعلاقات	تُستخلص الأمثلة باستخدام مجموعة صغيرة من قواعد استخراج العلاقات ومن ثمّ يُحتفظ بأبرزها، مع تعلم المزيد من القواعد والأمثلة بشكل متكرر	سهولة إضافة قواعد جديدة، وإمكانية تزويد تلك القواعد من قبل المستخدم	في الغالب تدني إمكانية الاستدعاء و/ أو إجراء تنقيح يدوي لتحقيق دقة عالية
الاستناد إلى القواعد	نص غير مصنف ومخططات علاقات وقواعد ومعاجم جغرافية	علاقات	تُستخلص العلاقات باستخدام قواعد الاستخراج ومعاجم كيانات الأسماء	سهولة إضافة قواعد جديدة، وإمكانية تزويد تلك القواعد من قبل المستخدم	في الغالب تدني إمكانية الاستدعاء وضرورة بذل جهد كبير في التطوير
الإشراف المباشر	نص غير مصنف ومخططات علاقات	علاقات	تدريب نموذج باستخدام مخطط علاقات وبيانات تدريب مصنفة	تعد هذه المنهجية حاليًّا الأعلى دقة وقدرة على الاستدعاء عندما يتعلق الأمر بعمليات استخراج العلاقات لمخطط معين	ضرورة بذل جهد مسبق في تصنيف البيانات ووجود خطر الإفراد في تجهيز طقم التدريب

المنهجية	المدخلات	المخرجات	الوصف	المزايا	العيوب
استخراج المعلومات المفتوح	نص غير مصنف	مجموعات علاقات	اكتشاف مجموعات العلاقات في النص باستخدام أسلوب التجميع، مع الاحتفاظ بأبرزها	لا داعي للمعرفة بالنص	صعوبة فهم معنى المجموعات وصعوبة تحويلها لمخططات علاقات
الإشراف عن بُعد	نص غير مصنف ومخططات علاقات وأمثلة	نموذج استخراج وعلاقات	تحشية بيانات التدريب آليًّا وتدريب نموذج بهدف استخراج المزيد من العلاقات، وذلك باستخدام مخطط علاقات وأمثلة علاقات	استخراج العلاقات عالية الاستدعاء والدقة	ضرورة وجود أمثلة أولية
المخططات الشاملة	عدة قواعد معرفة معبأة جزئيًّا	معرفة موحدة	أخذ عدد من قواعد المعرفة معرّفة بواسطة مخططات مختلفة ومعبأة جزئيًّا بالعلاقات، ومن ثمّ توقع صيغة موحدة لقواعد المعرفة	دمج العلاقات المعرّفة بواسطة مخططات مختلفة بعد عملية الاستخراج	عندما تكون قواعد المعرفة صغيرة، الأسرع إجراء هذه العملية يدويًّا

الفصل الخامس
ربط الكيانات

بمعرفتنا أيًّا من التعبيرات في النص تمثل الكيانات، تتلخص المهمة التالية في ربط الكيانات (أو إزالة الغموض في الكيانات) [111]، وعادةً يتطلب ذلك إضافة التعليقات على كيان يُحتمل أن يكون به بعض الغموض في مستند ما (على سبيل المثال: باريس) تحتوي على رابط إلى مُعرِّف مقبول يصف كيانًا فريدًا في إحدى قواعد البيانات أو علم الوجود (على سبيل المثال: http: / /dbpedia.org /resource /Paris). استخدمت منهجيات قواعد بيانات مختلفة الكيانات كهدف لإزالة الغموض (على سبيل المثال: صفحات ويكيبيديا [112-114]) وموارد البيانات المفتوحة المرتبطة (على سبيل المثال: DBpedia [115، 116]، YAGO [117]، Freebase [118]). العديد من أهداف التوضيح تتميز بالقواسم المشتركة والروابط، وفي معظم الأحيان يمكن الربط بينهما [119]، كما يعدُّ ربط إشارات الكيان بهذه الموارد أمرًا أساسيًّا للشروحات التلقائية الدلالية لوثائق الويب، وقواعد المعرفة، والبحث الدلالي، والوصول إلى المعلومات بمختلف اللغات، والمهام الأخرى ذات الصلة.

ربط الكيان مهمة صعبة للغاية، حيث تتطلب تلك الطرق معالجة تنوعات الاسم الأول، حيث يمكن الإشارة إلى الكيان نفسه بطرق مختلفة (مثل نيويورك والتفاحة الكبيرة)، بينما التحدي الثاني يمثل الغموض الكبير في الكيان، أي أن السلسلة نفسها ربما تشير إلى أكثر من كيان واحد (مثل باريس، فرنسا مقابل باريس، تكساس مقابل باريس هيلتون)، وبينما DBpedia يحتوي على ملايين الاحتمالات، يمثل غموض الكيان تحديًا صعبًا للغاية، حيث قد يكون للنص أكثر من مائة نتيجة في قاعدة المعرفة، وهناك تحد آخر صعب للغاية وهو وجود كيانات مفقودة، أي تكون النتيجة عدم وجود كيان مُستهدَف مناسب في قاعدة المعرفة.

تتضمن منهجيات ربط كيانات الأسماء NEL عادةً مرحلة اختيار المرشح، التي تحدد كافة مُدخلات قاعدة المعرفة المُرشَّحَة للكيان المُحدد المذكور في النص، ويلي ذلك مرحلة إزالة الغموض في المرجعية (أو تحليل الكيان)، التي تحدد الكيان المُستهدَف الأعلى احتمالاً بين جميع الكيانات المُرشَّحَة. تميل خطوة إزالة الغموض في المرجعية هذه إلى استخدام المعلومات السياقية من النص، وكذلك المعرفة من علم الأنماط لاختيار

عنوان URI المناسب. يمكن إزالة الغموض في الإشارات النصية إما بصورة منفصلٍ بعضها عن بعض، أو بصورة جماعية عبر الوثيقة بأكملها [116، 120].

الكثير من العمل حول ربط الكيانات يحقق فرضية العالم المغلق، أي أن هناك دومًا كيانا مُستهدَفا في قاعدة المعرفة، ومع ذلك، فالأمر بالنسبة للكثير من أنواع الوثائق (ولا سيما وسائل الإعلام الاجتماعية) وكذلك التطبيقات محدود للغاية، لأن تلك الكيانات عادةً تكون غير جديرة بالاهتمام، أو مُكتملة الأركان بشكل يمنع إدراجها في موسوعة ويكيبيديا أو مورد البيانات المفتوحة المُرتبطة LOD (يمكنك الرجوع إلى المناقشة السابقة في الفصل الثالث حول الكيانات الناشئة حديثًا)، ولذلك، فإن مهمة ربط كيانات الأسماء NEL الأكثر صعوبة هي إما إظهار نتيجة مُدخل مطابق من قاعدة المعرفة المُستهدَفة (على سبيل المثال:عنوان URI لـDBpedia، أو عنوان URL لويكيبيديا) أو NIL للإشارة إلى أنه لا يوجد كيان مطابق.

٥-١ ربط كيانات الأسماء والربط الدلالي

يهتم الربط الدلالي بمسألة كبيرة تتمثل في تحديد الموضوعات (مثل التكنولوجيا) والكيانات (على سبيل المثال: آي باد) التي تستحوذ على أفضل معنى للمستند. يُشار كذلك إلى الربط الدلالي بمهمة «aboutness» [121]، أو "C2W" (مفاهيم ويكيبيديا) ومهام "Sc2W" (مفاهيم مُسجلة في ويكيبيديا) [121].

عادةً يستند الربط الدلالي السليم إلى أدلة سياقية خفية، ويحتاج إلى الجمع مع المعرفة العالمية. على سبيل المثال، التغريدة التي يُذكر فيها آي باد تجعل شركة آبل كيانًا ذات صلة، وذلك بسبب العلاقة الضمنية بين الكيانين (آي باد وآبل)، مما يترتب عليه ألا يستلزم ذكر الكيانات والموضوعات المرتبطة بشكل صريح في نص الوثيقة، بينما من منظور تنفيذي، تشتمل مهمة الحيثية على تحديد الكيانات ذات الصلة على مستوى الوثيقة بأكملها، مع تخطي خطوة تحديد كيانات الأسماء NER التي تشتمل على تحديد إشارات الكيان الصريح أولاً.

على النقيض، فإن مهمة ربط كيانات الأسماء NEL المعنية في هذا الفصل، تتعلق بإزالة الغموض في الكيانات المذكورة صراحة فقط، وفي هذه الحالة، لا يلزم تحديد إشارات الكيان فقط من خلال تحديد وتصنيف كيانات الأسماء NERC، بل كذلك تحديد هوية الكيان الفريد المُستهدَف (أو لاشيء NIL) لُمعرفَات الكيان، وبما أن إشارات الكيان غير المُحددة لن يتم حذفها، فإن أداء ربط كيانات الأسماء NEL يعتمد بشكل كبير على أداء تحديد وتصنيف كيانات الأسماء NERC.

٥-٢ مجموعات البيانات لربط كيانات الأسماء NEL

تم إنشاء أول بنية لربط كيانات الأسماء NEL كجزء من مبادرات ربط الكيانات TAC-KBP [123، 124]، التي تحتوي على وثائق وكيان واحد محدد لكل وثيقة، وهو ما ينبغي توضيح ما إذا كان مدخلاً لقاعدة المعرفة أو لا شيئًا NIL، وفي حالة أن الكيان المذكور متوافر بالفعل، وهناك وثيقة واحدة فقط لكل وثيقة، فإن هذه البنية محدودة إلى حد كبير.

توجد قاعدة بيانات أقدم تدعى AQUAINT[1]، تحتوي على تعليقات وشروحات من نسخة قديمة من موسوعة ويكيبيديا، كما أنها ليست مخصصة لربط الكيانات المُعرفَة فقط بل تشمل مصطلحات من صفحات ويكيبيديا، مما يجعلها أكثر ملاءمة لتقييم الربط الدلالي، بدلاً من منهجيات ربط كيانات الأسماء NEL المستندة إلى البيانات المفتوحة المُرتبطة LOD.

تتكون بنية CoNLL/ AIDA [116] من مقالات إخبارية مشروحة مع مُعرِّفات الموارد المُوحَّدة YAGO وتنقسم إلى التدريب، والتطوير، والاختبار. تحتوي وحدة الاختبار وحدها على ٢٣١ وثيقة مع ٤,٤٨٥ من الشروحات المُستهدَفة.

سعيًا لمتابعة العمل، أصدر المؤلفون قاعدة بيانات أصغر AIDA-EE [125]، تحتوي على ٣٠٠ وثيقةٍ مع أسماء ٩,٩٧٦ كيانًا، مرتبطة بالإصدار ٢٠١٠ من موسوعة ويكيبيديا. هذه المجموعة من البيانات متحيزة نظرًا لأن كافة إشارات الكيانات تم

1- http://aksw.org/Projects/N3NERNEDNIF.html

التعرف عليها تلقائيًّا للمرة الأولى باستخدام أداة تحديد كيانات الأسماء ستانفورد NER، وتم ربط تلك الإشارات يدويًّا إلى صفحة ويكيبيديا المناسبة. بشكل عملي، هذا يعني أن إشارات الكيانات التي لم يحددها نظام ستانفورد سوف تعدُّ غير صحيحة أثناء التقييم، على الرغم من أن نظام ربط كيانات الأسماء NEL قد يكون صحيحًا.

هناك مجموعة بيانات حديثة أخرى هي N3[1]، تحتوي على ثلاثة مكانز باللغتين الإنجليزية والألمانية مع كيانات أضيفت إليها الحواشي والتعليقات يدويًّا، وهي مرتبطة بعناوين مُعرِّفات الموارد المُوحَّدة DBpedia URIs.

المكانز متناهية الصغر التي أنشئت خصيصًا لربط كيانات الأسماء NEL والتي تستند إلى البيانات المفتوحة المُرتبطة LOD تعد محدودة للغاية، على سبيل المثال، مكنز Ritter's [126]، يحتوي فقط على أنواع الكيانات، في حين أن تلك الكيانات من منافسات MSM [127، 128] جعلت إشارات اسم المستخدم وكذلك عناوين URL مجهولة المصدر. المكانز التي أنشئت للربط الدلالي، مثل Meij [121]، ليست مناسبة تمامًا لتقييم ربط كيانات الأسماء، نتيجة وجود كيانات ضمنية وموضوعات عامة (مثل «الموقع الإلكتروني»، «قابلية الاستخدام»، «الجمهور المستهدف»).

يحتوي مكنز YODIE الخاص بموقع تويتر على قرابة ٨٠٠ تغريدة، أضيف إليها التعليقات والحواشي بواسطة عنوان URI من DBpedia بواسطة العديد من الخبراء [129]. تحتوي التغريدات على وسوم وعناوين URLs وإشارات المستخدمين، بما في ذلك العديد من عناوين URIs من DBpedia المقابلة (على سبيل المثال: @eonenergyuk)، بينما تنقسم مجموعة البيانات[2] المتاحة بشكل عام إلى أجزاء تدريبية وتقييمية متكافئة.

1- https://gate.ac.uk/applications/yodie.html

2- https://gate.ac.uk/applications/yodie.html

٥-٣ المنهجيات المستندة إلى البيانات المفتوحة المُرتبطة LOD

عادة تحتوي طرق ربط الكيانات المستندة إلى علم الوجود وطرق إزالة الغموض على قاموس للمصطلحات لعنوان URI لكل كيان على حدة باستخدام صفحات كيانات ويكيبيديا، وعمليات إعادة التوجيه (المستخدمة للمرادفات والاختصارات)، وصفحات إزالة الغموض (لمختلف الكيانات التي تحمل الاسم نفسه)، والارتباطات التشعبية المستخدمة عند الربط بإحدى صفحات موسوعة ويكيبيديا. يستخدم هذا القاموس لتعريف جميع مُعرِّفات الموارد المُوحَّدة URIs لكيان مُعرِّف إحدى النصوص، وفيما يلي مرحلة إزالة الغموض، حيث يتم ترتيب جميع مُعرِّفات الموارد المُوحَّدة URIs المرشحة، وكذلك تحديد درجة الموثوقية. إن لم يكن هناك كيان مطابق في قاعدة المعرفة المُستهدَفة، تكون النتيجة هي NIL.

تستند الطرق النموذجية إلى إحصائيات مكنز ويكيبيديا إلى جانب التقنيات (على سبيل المثال: تردد المصطلح/ حجم الوثيقة TF /IDF) التي تتطابق مع المعرف الغامض في النص مقابل صفحات ويكيبيديا لكل كيان مرشح [115]. (ميشيلسون وآخرون) أوضحوا [130] كيف يمكن استخدام هذه المنهجية لاستخلاص الملف الشخصي الموضوع للمستخدم من تغريداته، استنادًا إلى التصنيفات المختلفة في موسوعة ويكيبيديا.

٥-٣-١ SPOTLIGHT DBPEDIA

واحد من أنظمة الشرح الدلالي المستندة إلى DBpedia المستخدمة على نطاق واسع هو DBpedia Spotlight [115]، وهو نظام مجاني قائم وقابل للتخصيص ومتوفر على شبكة الإنترنت، يشرح المستندات النصية من خلال عنوانات URIs من DBpedia، وهو يستهدف أنطولوجيا DBpedia، والتي تتميز بأكثر من ٣٠ فئة من المستوى الأعلى وإجمالي ٢٧٢ فئة. من الممكن تحديد الفئات (والفئات الفرعية المندرجة) المستخدمة للتعرف على الكيانات المُعرفَة، سواءٌ أكان بإدراجها صراحة أم من خلال استعلام SPARQL. تختار الخوارزمية في البداية الكيانات المرشحة عن طريق البحث في القاموس المستند إلى الموسوعة من ويكيبيديا الذي يحتوي على التعبيرات المفرداتية

لعنوانات URI، تليها مرحلة ترتيب عنوانات URI باستخدام نموذج الفضاء المتجه. ويرتبط كل مورد DBpedia بوثيقة، أنشئت من جميع الفقرات المذكور فيها هذا المفهوم في ويكيبيديا، ويتضح أن هذه الطريقة تتفوق على أداء OpenCalais و Zemanta (انظر القسم ٥-٤) بناء على اعتبار معيار ذهبي مصغر للمقالات الصحفية [115].

RT @XXXX Eyeopener vs. Ryerson Quidditch team this Sunday at 4 p.m. Anyone know where to get cheap brooms? #Ryerson @XXXX #Rams

@XXXX http://www.youtube.com/watch?v=eLMui7zBiXo we beat kilkenny after they beat us for the last 4 years in the hurling. Woo!!!

Kk its 22:48 friday nyt :D really tired so imma go to sleep :) good nyt x god bles xxxx

Amazon U.K. Offering HTC Desire Z Unlocked http://dbpedia.org/resource/Irish_Museum_of_Modern_Art earlier in Lo... http://bit.ly/bsyz9H URL

RT @XXXX: Eventful morning for Oklahoma State's Darrell Williams. Won Big 12 Rookie of the Week Award-and got charged with f...

الشكل ٥-١: نتائج DBpedia Spotlight حول التغريدات.

يبين الشكل ٥-١ العديد من التغريدات التي أضيفت لها التعليقات والشروحات في DBpedia Spotlight، حيث تُظهِر النتائج بوضوح الحاجة إلى التدقيق الإملائي للتغريدات، وكذلك الصعوبات التي واجهت Spotlight في تمييز عنوانات URLs، وكما يتضح هنا، صُمِّمَت الخوارزمية بشكل افتراضي لتوسيع الاستدعاء (أي إضافة التعليقات والشروحات إلى أكبر عدد ممكن من الكيانات، باستخدام الملايين من الحالات من DBpedia). نظرًا للطبيعة القصيرة والصاخبة للتغريدات، حيث من الممكن أن يؤدي ذلك إلى نتائج غير دقيقة، مما يترتب عليه حتمية إجراء مزيد من التقييم الرسمي المستند إلى مجموعة كبيرة من البيانات المشتركة من الرسائل القصيرة في وسائل التواصل الاجتماعية، لتحديد أفضل القيم لمختلف معاملات DBpedia Spotlight (على سبيل المثال: الموثوقية، والدعم).

٥-٣-٢ YODIE

إطار إزالة غموض الكيانات المستندة إلى مورد البيانات المفتوحة المُرتبطة LOD

YODIE[1] هو إطار NED مُستند إلى GATE، وهو يجمع بين نظام ANNIE NER من GATE وعدد من استراتيجيات اختيار مرشح مُحدِّد المصادر المُوحَّد URI المستخدمة على نطاق واسع، ومقاييس التشابه، ونموذج التعلم الآلي لإزالة الغموض عن الكيان، الذي يحدد أفضل مُحدِّد مصادر مُوحَّد URI مرشح. لكل إشارة NE ولكل مرشح، يقوم إطار YODIE بحساب عدد من الدرجات القياسية المنتظمة التي تعكس التشابه الدلالي بين الكيان المشار إليه من قبل المرشح وسياق الإشارة الخاص به:

- نتائج الارتباط: أدخلت في [131]، وتستخدم نسبة الروابط الواردة التي تتداخل في مخطط ويكيبيديا البياني لإعطاء أفضلية إلى خيارات المرشحين المتطابقة.
- التشابه المستند إلى مورد البيانات المفتوحة المُرتبطة LOD: يشبه الموضح أعلاه، ولكنه يستند إلى عدد العلاقات بين كل زوج من عنوانات مُحدِّد المصادر المُوحَّد URIs في الرسم البياني DBpedia (موضح فيما يلي).
- نتائج التشابه المستندة إلى النصوص: تقيس هذه النتائج مدى التشابه بين السياق النصي لكيانات الأسماء والنص المقترن بكل عنوان مُحدِّد المصادر المُوحَّد URI الخاص بهذه الإشارة (انظر أدناه).

عملية تحديد كيفية الجمع بين هذه النتائج لاختيار أفضل مُحدِّد مصادر مُوحَّد URI هذه العملية ذات أهمية كبيرة، ويستخدم YODIE LibSVM[2] لتحديد أفضل مرشح.

تتكون بيانات التدريب الخاصة بالنموذج من حالة تدريبية واحدة لكل مرشح يقوم بإنشائها النظام في بنية التدريب، حيث تحصل كل حالة على هدف صحيح إذا كان المرشح هو الهدف الصحيح لإزالة الغموض، بينما تحصل كل حالة على هدف خاطئ إذا

1- http://www.csie.ntu.edu.tw/-cjlin/libsvm/

2- http://www.nist.gov/tac/2013/KBP/

حصل خلاف ذلك. تستخدم مختلف قيم مقاييس التشابه كخصائص للمقارنة، مما يعني أنه في وقت التطبيق، يعين النموذج لكل مرشح حالة إما صحيحة أو خاطئة، إلى جانب واحدة من الاحتمالات. عملية التصنيف هذه تجري بصورة مستقلة عن المرشحين الآخرين لذلك الكيان، ولكن يمكن ترتيب قائمة المرشحين استنادًا إلى الاحتمالات، ولذا يتم تعيين مُحدِّد المصادر المُوحَّد URI الأكثر احتمالاً بينما يتم إزالة الغموض عن هذا الكيان، ما لم تكن الاحتمالية الخاصة بهذا الكيان أقل من درجة محددة، وفي هذه الحالة يتم تعيين «NIL». بيانات التدريب لهذا النموذج تستند إلى بيانات TAC KBP في الفترة بين ٢٠٠٩ و٢٠١٣، باستثناء مجموعة[1] ٢٠١٠، إلى جانب مجموعة التدريب AIDA [116]، ومجموعة تدريب التغريدات المُشار إليها في القسم ٥-٢.

٥-٣-٣ مناهج رئيسة أخرى مستندة إلى مورد البيانات المفتوحة المُرتبطة LOD

هناك اثنان من الأنظمة الأخرى المتوافرة، من نوعي أنظمة NED المستندة إلى مورد البيانات المفتوحة المُرتبطة LOD، وهما نظام AIDA [116، 125] ونظام AGDISTIS [120]، وكلاهما منهجان أساسهما إزالة الغموض المُستند إلى الرسوم البيانية، ويهدفان معًا إلى إزالة الغموض في جميع الكيانات المذكورة في النص. في حين أن هذه المناهج تميل إلى العمل بشكل رائع في الوثائق كبيرة الحجم، يكون أداؤها في التغريدات وغيرها من منشورات وسائل التواصل الاجتماعية القصيرة سيئًا إلى حد كبير.

AGDISTIS [120] هو منهج ربط كيانات الأسماء NEL المستند إلى الرسوم البيانية المُصمم ليكون بمنزلة أداة تشخيص قاعدة المعرفة، فهو يجمع بين خوارزمية البحث الموضوعي المُستحدث من النص التشعبي (HITS) إلى جانب استراتيجيات توسعة التسمية وعوامل تشابه الارتباط. تم اختبار المنهج باستخدام كل من DBpedia وYAGO2، وعلى غرار معظم أنظمة ربط كيانات الأسماء NEL الأخرى الموضحة هنا، يقوم بإزالة الغموض المتعلق بالتصنيفات الثلاثة القياسية؛ الشخص، والمنظمة، والمكان. في البداية، بالنسبة لكل كيان من كيانات الأسماء، يتم تحديد عدد من المرشحين،

1- http://wikipedia-miner.cms.waikato.ac.nz/

وفي الخطوة التالية، تستخدم خوارزمية HITS لحساب التخصيص الأمثل من خلال إنشاء رسم بياني لإزالة الغموض. تم اختيار جميع خوارزميات التعقيد المؤقتة متعددة الحدود فقط، لذلك ينطبق AGDISTIS على وثائق الويب كبيرة الحجم.

هناك مثال آخر TagMe، المُصمَّم خصيصًا لشرح النصوص القصيرة فيما يتعلق بالموسوعة ويكيبيديا [132]. هناك تقرير مُفصَّل حول التقييم المُقارَن للمنهجيات الحديثة العامة كافة، باستثناء المنهجية الأحدث من AGDISTIS، في [122]، وذلك باستخدام العديد من مجموعات البيانات الإخبارية المتوافرة.

في النهاية، فإن نظام ربط كيانات الأسماء NEL المرتبط بـ YAGO هو إطار LINDEN [117]، نظام NEL يستفيد من المعلومات الدلالية الأكثر ثراء في YAGO (التشابه الدلالي)، بالإضافة إلى المعلومات المستندة إلى ويكيبيديا (باستخدام بنية الارتباط للارتباطية الدلالية). تعتمد هذه الطريقة بشكل كبير على مجموعة أدوات مُنقِّب ويكيبيديا [114][1]، الذي يُستخدَم لتحليل سياق إشارة الكيان الغامض وتحديد مفاهيم ويكيبيديا. أظهر تقييم مجموعة بيانات TAC-KBP2009 تفوق LINDEN على أفضل الأنظمة المُستندة إلى ويكيبيديا فقط التي خضعت لتقييم TAC الأوليّ. لسوء الحظ، لم تتم مقارنة LINDEN مباشرة مع DBpedia Spotlight من حيث مجموعة بيانات التقييم المشتركة.

٥-٤ الخدمات التجارية لربط الكيانات

هناك عدد من خدمات ربط الكيانات التجارية على شبكة الإنترنت تقوم بتعيين عنوانات URIs الخاصة بالبيانات المرتبطة، أداة NERD على شبكة الإنترنت [119] تسمح بالمقارنة السهلة وفق مجموعات البيانات التي يقوم بتحميلها المستخدم، كما تقوم بتوحيد نتائجها ورسم العلاقات البيانية بينها إلى سحابة البيانات المرتبطة المفتوحة. سوف نركز هنا فقط على الخدمات التي تستخدمها أساليب البحث التي نستعرضها [133-135].

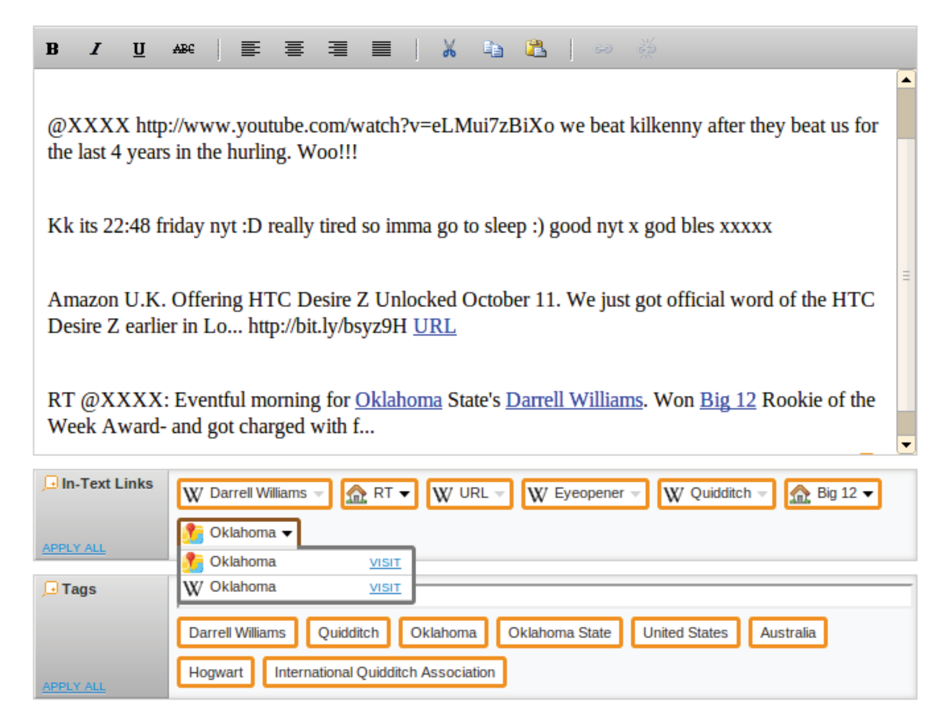
1- http://vvv.zamanta.com

Zemanta[١] هي أداة الشرح الدلالي على شبكة الإنترنت، التي طورت في الأساس لأنظمة المدونات ورسائل البريد الإلكتروني لمساعدة المستخدمين على إدراج الوسوم، والروابط من خلال التوصيات. يعرض الشكل ٥-٢ مثالاً للنص والعلامات الموصى بها، والأهداف المحتملة للروابط النصية (مثل مقالة ويكيبيديا W3C وصفحة W3C الرئيسة)، وغيرها من المقالات ذات الصلة، ومن ثم يعود الأمر إلى المستخدم ليقرر أيًّا من العلامات يجب استخدامها والأهداف المحتملة للروابط النصية التي يرغب في إضافتها. في هذا المثال، تم تظليل الروابط النصية الخاصة بالمصطلحات باللون البرتقالي، وكلها تشير إلى مقالات ويكيبيديا حول الموضوعات ذات الصلة.

OpenCalais هي إحدى الخدمات التجارية لإضافة التعليقات والشروحات الدلالية على شبكة الإنترنت، والتي تستخدم من قبل بعض الباحثين في مجال وسائل التواصل الاجتماعية. على سبيل المثال، (أبيل وآخرون). [134] استخدموا OpenCalais للتعرف على كيانات الأسماء في التغريدات الإخبارية[٢]. الكيانات المُستهدَفة عادة ما تكون المواقع والشركات والأشخاص والعنوانات وأرقام الهاتف والمنتجات والأفلام، ...الخ. الأحداث والحقائق التي يتم استخلاصها هي تلك التي تحتوي على الكيانات المذكورة أعلاه، على سبيل المثال، الاستحواذات، والتحالفات التجارية، والشركات المنافسة. يبين الشكل ١, ٨ مثالاً على نص أضيفت له التعليقات والحواشي باستخدام بعض الكيانات.

١ - للأسف، لم يقوموا بتقييم مدى دقة تعريف كيانات الأسماء من OpenCalais في مجموعة البيانات الخاصة بهم.

2- http://www.nlm.nih.gov/research/umls/



الشكل ٥-٢: واجهة وسوم Zemanta على شبكة الإنترنت.

تحتوي التعليقات التوضيحية للكيانات على عناوين URIs التي تسمح بالدخول عبر HTTP للحصول على معلومات إضافية حول هذا الكيان عبر البيانات المرتبطة. في الوقت الحالي، ترتبط وصلات OpenCalais بثمانية من مجموعات البيانات المرتبطة، بما في ذلك قاعدة المعرفة الخاصة بها، وDBpedia، وويكيبيديا، وIMDB، وShopping.com. هذه الأمثلة تتوافق بشكل عام مع أنواع الكيانات المندرجة تحت علم (الأنطولوجيا).

القيد الرئيس لخدمة Calais تتمثل في طبيعته الاستحواذية، ولتوضيح ذلك، يقوم المستخدمون بإرسال المستندات التي سوف يضاف إليها التعليقات والشروحات بواسطة خدمات الويب، ويتلقّون النتائج لاحقا. ولكن لا تتوفر لهم الوسيلة لإعطاء Calais وجودية مختلفة لإضافة التعليقات والحواشي أو لتخصيص الطريقة التي تعمل من خلالها طبيعة استخراج الكيان.

٥-٥ ربط كيانات الأسماء NEL لمحتوى وسائل التواصل الاجتماعية

طُوِّرت منهجيات ربط كيانات الأسماء NEL المستندة إلى البيانات المفتوحة المُرتبطة LOD والتي تعدُّ أحدث التقنيات في هذا المجال وتمت مناقشتها سابقا وتم تقييمها استنادًا إلى المقالات الإخبارية وغيرها من النصوص المكتوبة بعناية، والنصوص الطويلة [111، 122]، وفي القسم ٥-٢ أوضحنا أنه يوجد عدد قليل للغاية من بنية المدونات الصغيرة المشروحة من خلال عنوانات URIs المُستندة إلى البيانات المفتوحة المُرتبطة LOD وهي بالإضافة إلى ذلك صغيرة وغير مكتملة.

علاوة على ذلك، قام الباحثون بتقييم ربط كيانات الأسماء NEL للمدونات الصغيرة، على سبيل المثال، [67]، أوضحت المنهجيات المتطورة نوعًا من الأداء الضعيف، نظرًا للسياق المحدود، والتشويشات اللغوية، واستخدام الرموز التعبيرية، والمختصرات، والوسوم. يتم التعامل مع كل منشور في المدونات الصغيرة بشكل منفصل، دون الأخذ بعين الاعتبار السياق الأعرض نطاقًا، وبشكل خاص، تتم معالجة نصوص التغريدة فقط، على الرغم من حقيقة أن كائن JSON خاص بالتغريدة يحتوي أيضًا بيانات الملف الشخصي لصاحب التغريدة (الاسم بالكامل، والموقع الاختياري، ونصوص الملف الشخصي، وصفحة الويب). تقريبًا ٢٦٪ من جميع التغريدات تحتوي كذلك على عنوانات URLs [136]، و٦,١٦٪ من الوسوم، و٨,٥٤٪ من واحد على الأقل من إشارات المستخدم.

ربط كيانات الأسماء للمدونات الصغيرة تعد مهمة حديثة نسبيًّا، وبها الكثير من الأمور التي لم تُكتشف بعد، حيث أظهرت التقييمات المؤخرة التي تركز على التغريدات للمرة الأولى مشكلات في استخدام أحدث منهجيات ربط كيانات الأسماء NEL في هذا الصدد [67، 134]، ويرجع ذلك إلى حد كبير إلى إيجاز التغريدات (١٤٠ حرفًا). ليس هناك الكثير من الأبحاث حول تحليل وسوم تويتر وشرحها من خلال مُدخلات DBpedia، لتعزيز البحث الدلالي حول محتوى المدونات الصغيرة، في [137] مثالاً على ذلك. بينما حققت منهجيات تستند إلى الرسوم البيانية المعرفية للتغلب على التحديات المتمثلة في وجود سياق محدود جدًّا بعض النجاح في هذا الصدد [138].

استخدم شين وآخرون [139] مزيدًا من التغريدات من المنشورات اليومية للمستخدم لتحديد الموضوعات المُحددة لهوية المستخدم واستخدامها لتحسين إزالة الغموض. (هوانغ وآخرون) [140] قاموا بعمل امتداد لإزالة الغموض المستند إلى الرسم البياني حيث يعرض «مسارات فوقية» توضح السياق من تغريدات أخرى من خلال الوسوم المشتركة، وصاحب التغريدات، أو الإشارات.

غاطاني وآخرون [141] استفادوا من توسيع عنوان URL واستخدموا السياق المستمد من تغريدات المستخدم نفسه التي تحتوي على الوسوم نفسها، ولكن لم يقيِّموا مساهمة هذا السياق في الأداء النهائي، وكذلك لم يستفيدوا من مُعرفَات الوسوم أو الملفات الشخصية للمستخدم.

أحد الأبحاث الأخيرة [129] درس التأثير على أداء ربط كيانات الأسماء NEL لاستخدام توسعة السياق، والمعلومات حول السيرة الذاتية للمستخدم، ومُعرِّفَات الوسوم، وبشكل خاص، في حالة الوسوم، يتم إثراء محتوى التغريدات باستخدام مُعرفَات الوسوم، التي يتم استردادها تلقائيًّا من شبكة الإنترنت. وكذلك، يتم إثراء التغريدات التي تحتوي على الإشارة @mentions بالمعلومات النصية من الملف الشخصي على تويتر. في حالة عنوانات URLs، يتم إلحاق محتوى الويب المقابل إلى التغريدة، بينما يُقاس أداء إزالة الغموض سواءٌ أكان عند تنفيذ هذا التوسع في السياق بشكل فردي (أي الوسوم فقط، وعنوانات URLs فقط، ...الخ)، أم عند استخدام الأنواع الثلاثة من المعلومات السياقية معًا.

٥-٦ المناقشة

أثبت استعراض ربط الكيانات المستندة إلى موسوعة ويكيبيديا والمستندة إلى البيانات المفتوحة المُرتبطة LOD أن غالبية الدراسات ركزت على عدد قليل من الكيانات الشائعة، والمفهومة جيدًا؛ وتحديدا الأشخاص، والمواقع، والمنظمات، وفي بعض الأحيان المنتجات. تتمحور التحديات الحقيقية في توسعة هذه المجموعة لتشمل أنواعًا جديدة من الكيانات، حيث سيؤدي ذلك أيضًا إلى زيادة الغموض، ومن ثم إلى

الحد من أداء أساليب ربط كيانات الأسماء NEL، وثمة مشكلة رئيسة أخرى لم تدرس بالشكل الوافي حتى الآن وتتمثل في تحسين خوارزميات ربط كيانات الأسماء NEL لمنشورات وسائل التواصل الاجتماعية، حيث يكون السياق والمحتوى النصي مختلفين تمامًا، مما يجعل من الصعب معالجتها بدقة.

التحدي الرئيس الآخر يتمثل في توسعة النطاق ليشمل لغات أخرى غير اللغة الإنجليزية، حيث يحتاج الباحثون كذلك إلى مجموعات بيانات جديدة من التدريب والتقييم، وخاصة تلك التي تتعلق بمحتوى وسائل التواصل الاجتماعي، في حين أن هناك بعض الطرق التي تعالج العديد من اللغات (على سبيل المثال: DBpedia Spotlight، وYODIE)، بينما لا يزال الجزء الأكبر من الأبحاث حول ربط كيانات الأسماء NEL يجري على مجموعات بيانات اللغة الإنجليزية.

الفصل السادس
تطوير الأنطولوجيا الآلي

٦-١ مقدمة

في هذا الفصل، سوف نستعرض مفهوم تطوير الأنطولوجيا [أو كما يُطلق عليها «خرائط المعاني أو المفاهيم»] بصورة آلية والذي يضم ثلاثة مكونات، وهي التعلم والتعبئة والتنقيح. تشير عملية التعلم الأنطولوجي (التوليد الأنطولوجي) إلى مهمة إنشاء أنطولوجيا جديدة بدءًا من الصفر، وتتعلق بصفة عامة بمهمة تحديد المفاهيم وتوليد العلاقات ذات الصلة بين تلك المفاهيم. تتكون عملية تعبئة الأنطولوجيا من إضافة الحالات (instances) إلى هيكل أنطولوجي موجود مسبقًا (جرى إنشاؤه على سبيل المثال بواسطة مهمة التعلم الأنطولوجي). تشمل مهمة تنقيح الأنطولوجيا إضافة مفاهيم وعلاقات و/أو حالات (instances) جديدة أو حذفها أو تغييرها ضمن أنطولوجيا موجودة مسبقًا. يمكن استخدام التعلم الأنطولوجي أيضًا للإشارة إلى جميع المهام الثلاث، وبالأخص عندما يتم تنفيذ مهمتي التعلم والتعبئة عبر منهجية واحدة. تتمثل نقطة البداية عادة في جميع مكونات عملية تطوير الأنطولوجيا بمكنز كبير يضم نصوصًا غير مهيكلة (قد يكون هذا المكنز شبكة الإنترنت بأكملها، أو مجموعة من الوثائق ذات نطاق حر). نحن لسنا مهتمين هنا بعملية إنشاء الأنطولوجيا بدءًا من الصفر، لأنها لا تشمل في العادة استخدام أساليب معالجة اللغات الطبيعية.

في بقية أجزاء هذا الفصل، سوف نشرح هذه المهمة بالتفصيل، كما سنشرح ما تحمله من أوجه شبه واختلاف مع عملية إضافة التعليقات والشروحات (annotation)، وسنقدم أمثلة تدلّ على فائدتها. بعد ذلك سوف نشرح عددًا من المنهجيات المعتادة، ومرة أخرى سنبني على أساس الأدوات التي ورد شرحها في الفصول السابقة. ينبغي ملاحظة أن هناك عددًا من الكتب المهمة التي تتناول تعلم الأنطولوجيا وتعبئتها، وتختلف هذه الكتب في المنظور الذي اعتُمد عليه في تأليفها -راجع، على سبيل المثال [142–144]. إذًا سنقدم في هذا الفصل تلخيصًا لعدد من أبرز المفاهيم، وذلك من منظور معالجة اللغات الطبيعية.

٦-٢ المبادئ الأساسية

من الواضح أن الأنطولوجيات ذات أهمية قصوى في تطبيقات الويب الدلالي. وفي حين يوجد الآلاف من الأنطولوجيات الموجودة مسبقًا، التي تتراوح في حجمها ما بين أنطولوجيات ذات نطاق صغير - وذات تطبيق محدد، إلى أنطولوجيات ضخمة وشاملة مثل DBpedia، إلا أنها عادة ما تكون غير كافية أو غير مناسبة لمهمة معينة. أضف إلى ذلك أن الأدوات والتطبيقات الجديدة قد تتطلب أنواعًا جديدة من الأنطولوجيات، على سبيل المثال، يتطلب الاهتمام المتزايد في الآونة الأخيرة تعدين الآراء داخل تقييمات المنتجات أنطولوجيات خاصة قادرة على التعرف على خصائص معينة في المنتجات. إذا كان المرء يرغب في تحليل الآراء المتعلقة بالكاميرات، فعليه معرفة جميع المكونات المختلفة للكاميرا وطبيعة العلاقة بينها - العدسات وأنواع البطاريات والمقاسات والجهة المصنعة وما شابهها. وبالمثل، تضم الفنادق خصائص من قبيل عدد الغرف والمطعم والمقهى وحمام السباحة والخدمة وغيرها. هذه الخصائص ليست من مكونات الفندق بالمعنى الدقيق للكلمة، لذا فإنها قد لا ترد بالضرورة في «أنطولوجيا فندق» نموذجية. سوف نتناول تعدين الآراء المتصل بالخصائص بشكل أكبر في الفصل السابع.

بصورة عامة، ليست عملية إنشاء الأنطولوجيات يدويًّا مجدية أو قابلة للتطبيق، ما عدا الأنطولوجيات الخاصة بنطاقات محدودة جدًّا كلعب الأطفال، أو في حالات خاصة جدًّا، وهي تتطلب جهدًا بشريًّا وتكاليفَ كبيرة، إلى جانب كونها غير موضوعية. من جهة أخرى، فإن الإنشاء الآلي للأنطولوجيات معرض للأخطاء، فجودته رهن بجودة البيانات التي يتم توليد الأنطولوجيا منها في أحسن الحالات، ونادرًا ما تكون هذه البيانات كاملة، كما أنها تسبب معضلة من ناحية أن استخراج العلاقات الصحيحة بين عناصر الأنطولوجيا ليست بالمهمة السهلة، وذلك نظرًا لأن هذه المعلومات نادرًا ما ترد صراحة في البيانات. لذا يجب السعي لإيجاد حل وسط بين الإنشاء الآلي بالكامل للأنطولوجيا وتقليل الخسارة في الأداء لأدنى الحدود من جهة، والذاتية من جهة أخرى.

وفي حين توجد أنطولوجيات خاصة بنطاق معين، وفي بعض المجالات تكون هذه الأنطولوجيات شاملة (يوجد في المجال الطبي، على سبيل المثال، قواعد معرفة ضخمة

مثل قاعدة UMLS[1] المعرفية وأنطولوجيا الجينات[2])، لكن مع ذلك من غير المحتمل أن تكون أي قاعدة معرفة موجودة مسبقًا كافية تمامًا لأي تطبيق من تطبيقات الويب الدلالي. وإلى جانب احتمال احتوائها على أخطاء أو إغفال أو تكرار، فإنها قد تكون شديدة الغموض أيضًا. علاوة على ذلك، قد تتطلب الأنواع المختلفة من التطبيقات داخل النطاق نفسه أنواعًا مختلفة من الأنطولوجيات، فقد لا تكون أنطولوجيا طبية عامة محددة بما فيه الكفاية لأداء المهمة في نطاق فرعي مثل نطاق أمراض العيون مثلاً.

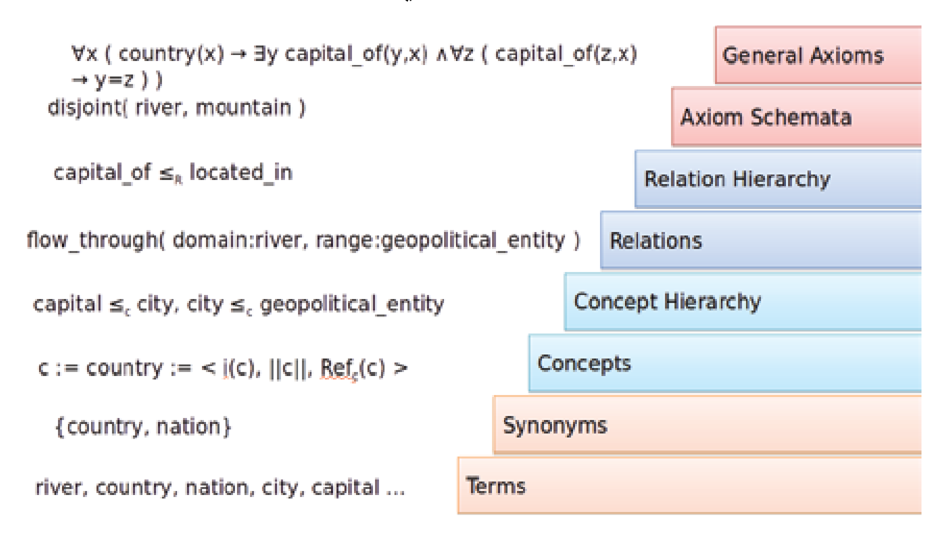
هناك مشكلة أخرى وهي عدم توحيد المصطلحات. وفي المقابل حتى عند توحيد المصطلحات، قد تظل أشكال مختلفة للمصطلحات قيد الاستخدام في مصادر النصوص، مثل تعبير نوبة قلبية أو تعبير احْتِشاءُ عَضَلِ القَلْب. تكون الكثير من المصطلحات على درجة عالية من الغموض أيضًا، وهذا لا يقتصر على التفاوت بين المصطلحات من نطاق إلى آخر (مثال: يختلف مصطلح فأرة في نطاق علم الحاسوب عنه في نطاق علم الحيوان)، بل يشمل أيضًا الغموض داخل النطاقات نفسها (عادة بسبب التدني في دقة التعبير، مثال، قد يشير مصطلح رِجل في الطب إلى الرجل البشرية أو الاصطناعية). زيادة على ذلك، قد يشير نصٌّ ما في نطاق معين إلى مفهوم يقع خارج ذلك النطاق ويحمل معنى يتداخل مع مفهوم يقع داخل ذلك النطاق (مثال: ورود الجملة التالية في تقرير طبي: ارتجاج في المخ نجم عن ضرب رأسها على رجل طاولة). ينبغي الأخذ في الاعتبار أساليب تكييف الأنطولوجيات مع المهمة والمجال من أجل تحقيق إمكاناتها بشكل كامل في التطبيقات. لذا تكون مهمة تخصيص المصادر المعجمية شديدة الأهمية، وهنا تلعب مهمتا التجميع وتمييز المصطلحات دورًا مهمًّا من خلال هيكلة المعرفة المطلوبة.

يمكن وصف العناصر والمنهجيات الأساسية التي تتكوّن منها عملية تطوير الأنطولوجيات بأنها تشبه كعكة طبقات التعلم الأنطولوجي (Layer Cake) (الشكل رقم ٦-١)، بناءً على فكرة كعكة طبقات الويب الدلالي (Semantic Web layer cake) المشهورة [145]. بدءًا من أسفل الكعكة وانتقالاً إلى أعلاها، تتمثل المهام الأساسية

1- http://geneontology.org/

2- http://code.google.com/p/jatetoolkit/

في تمييز المصطلحات والمترادفات، حيث يجوز أن تكون المصطلحات عبارة عن مدن وبلدان مثلاً. تضم المستويات التالية المفاهيم والأنواع والعلاقات (الخصائص)، على سبيل المثال، تنتمي المدن إلى البلدان، وبعض المدن عواصم، ويوجد في البلدان عواصم. أخيرًا، يوجد لدينا في القمة بديهيات (axioms) مثل الانفصال (disjointness) لا يمكن للشيء أن يكون نهرًا وجبلاً في الوقت نفسه). بالطبع هذه نظرة مبسطة نوعًا ما إلى الأمور، وفيها بعض القيود، وهي مبنية على اتباع منهجية معجمية لغرض الحصول على الأنطولوجيات [146]. غير أن هذه المنهجية هي بالذات المنهجية التي نتبعها في هذا الفصل، وذلك لأن محور اهتمامنا يدور حول أساليب معالجة اللغات الطبيعية المستخدمة لغرض تطوير الأنطولوجيات، لذا فهي مناسبة جدًّا.



الشكل ٦-١ : كعكة طبقات التعلم الأنطولوجي (مقتبس من جيميانو، ب.: تعلم الأنطولوجيات وتعبئتها من النص: الخوارزميات والتقييم والتطبيق، سبرينجر-فيرلاج، نيويورك، ٢٠٠٦).

٦-٣ استخراج المصطلحات

إن التعرف على المصطلحات ذات الصلة بالنطاق هي خطوة أولى مهمة في كل من مهمتي تعبئة الأنطولوجيات وتوليدها، وتُعرف هذه المهمة بمهمة استخراج أو تمييز المصطلحات، وتعرف اختصارًا بـATR (التعرف الآلي على المصطلحات). بوجه عام، تجري عملية تعبئة الأنطولوجيات آليًّا بواسطة نوع من أنواع أساليب استخراج

المعلومات المستندة إلى الأنطولوجيات (OBIE)، كما ورد شرحه في الفصل الخامس. وفي حين تتعلق مهمة استخراج المعلومات المستندة إلى الأنطولوجيات في العادة على تمييز كيانات الأسماء وربطها بإحدى الأنطولوجيات، وذلك لغرض تعبئة الأنطولوجيا، تتكون هذه المهمة من تحديد المصطلحات الرئيسة داخل النص ومن ثمّ ربطها بالمفاهيم الواردة في الأنطولوجيا (استخراج العلاقات). في مهمة توليد الأنطولوجيا، يُعثر أولاً على المصطلحات وبعد ذلك تُستخلص العلاقات الموجودة بينها، وهو ما يشكل أساس الأنطولوجيا نفسها.

يدور جدل كبير حول تعريف الـ«مصطلح». بصفة عامة، يمكن القول: إن المصطلح يشير إلى مفهوم محدد يحمل سمة من سمات نطاق أو لغة فرعية. وخلافًا لكيانات الأسماء كالأشخاص والمواقع التي عادة ما تكون ذات طبيعة عامة في مختلف النطاقات، إلا أن مصطلحًا تقنيًّا من قبيل احْتِشاءُ عَضَلِ القَلْب يصبح تعبيرًا ذا صلة فقط عندما يرد في أحد المجالات الطبية، لكن لو كنا مهتمين بالمصطلحات الرياضية، فلن يُنظر إليه على الأرجح على أنه تعبير ذو صلة، حتى لو ورد في مقال رياضي. وكما هو الحال مع كيانات الأسماء، تتشكل المصطلحات عمومًا من العبارات الاسمية. في بعض السياقات، ولا سيما في سياق الأنطولوجيات الموجودة مسبقًا، يمكن اعتبار الأفعال على أنها مصطلحات، لكن غالبية أساليب تمييز المصطلحات المستندة إلى المكانز لا تعتبرها كذلك. قد يختلف تعريف العبارة الاسمية نفسه من مكان لآخر، فكما شرحنا في الفصل الثاني، قد تقوم بعض أدوات تجزئة النص باستخراج عبارات اسمية تضم عبارات حروف الجر، وقد لا يقوم بعضها الآخر بذلك.

يمكن تنفيذ مهمة تمييز المصطلحات بعدة طرق. يتمثل وجه الاختلاف الأهم الذي نعرضه هنا في الاختلاف بين الخوارزميات التي لا تأخذ بعين الاعتبار سوى الخصائص التوزيعية للمصطلحات، مثل التكرار ومعامل تحديد الوزن tf/idf (تكرار المصطلح/ عكس تكرار المستند) [147]، وأساليب الاستخراج التي تستخدم المعلومات السياقية ذات الصلة بالمصطلحات. غير أن العديد من المنهجيات تجمع بين نوعي المعرفة. في العادة، تُستخدم الأساليب اللغوية في المقام الأول بغية إيجاد المصطلحات المحتملة، ومن ثمّ تُصنف هذه المصطلحات وفقًا لمدى أرجحية المصطلح. بعد ذلك يمكن

استخدام نقطة بداية (حد أدنى مقترح، بالإنجليزية threshold) لاتخاذ قرار مطلق بين ما يمكن اعتباره مصطلحًا وما لا يمكن اعتباره كذلك، وهذه خطوة شديدة الأهمية في معظم التطبيقات. بالنظر لكون مهمة تقييم عملية تصنيف المصطلحات وتمييزها بالغة الصعوبة وذاتية، حيث يمكن أن يختلف الحل الأمثل اعتمادًا على طبيعة المهمة، فقد جرى تطوير مجموعة من أطر العمل الخاصة باستخراج المصطلحات، حيث يمكن تجريب جميع الحلول أو الأشكال المختلفة ومقارنة بعضها ببعض. من الأمثلة الجيدة على ذلك نظام TermRaider (سيأتي شرحه) ونظام JATE[1].

٦-٣-١ منهجيات المعرفة التوزيعية

تستخدم هذه المنهجيات في العادة أساليب تعتمد على التكرار مبنية على أساس نموذج tf/idf. يعكس نموذج tf/idf (تكرار المصطلح/ عكس تكرار المستند) مدى أهمية الكلمة بالنسبة لمستند ما ضمن مجموعة. ونظرًا لورود بعض الكلمات بصورة متكررة جدًّا في جميع النطاقات، تصبح قيمة tf/idf معدلة تبعًا لذلك، حيث تزداد طرديًّا مع زيادة عدد مرات ورود كلمةٍ ما في المستند، لكن تكرار الكلمة في المكنز يوازن ذلك. يتمثل المبدأ الذي يستند عليه استخدام هذه القيمة في مهمة استخراج المصطلحات في أننا نتوقع أن ترد المصطلحات بتكرار أكبر في مكنز ما ذي صلة بالنطاق، أكثر من ورودها في نطاق غير ذي صلة، في حين أن غير المصطلحات (non-terms) سوف تظهر في كلا المكنزين موزعة بالتساوي، أو حتى بتكرار أقل في المكنز الخاص بالنطاق. على سبيل المثال، نتوقع أن يرد المصطلح احْتِشاءُ عَضَلِ القَلْب بتكرار أكبر في مكنز طبي مقارنة بمكنز مؤلف من النصوص الرياضية. إذًا، نستخدم نموذج tf/idf في العادة للمقارنة بين مكنز خاص بنطاق معين ومكنز عام، بدلاً من مقارنة مستند واحد بمكنز واحد.

هناك العديد من الاختلافات والتحسينات المدخلة على نموذج tf/idf الأساسي. نظام TermRaider[2] من الملحقات الإضافية ضمن منصة GATE المستخدمة في مهمة استخراج المصطلحات التي تقوم بتوليد المصطلحات المحتملة من أحد المكانز،

1- https://gate.ac.uk/projects/arcomem/TermRaider.html

2- http://www.nactem.ac.uk/software/termine/

إلى جانب درجة المصطلحية (statehood) المشتقة إحصائيًّا. ومثل معظم أساليب استخراج المصطلحات، يتعرف النظام أولاً على المصطلحات المحتملة بناءً على المبادئ اللغوية، وبعد ذلك يقوم بتصفيتها وتصنيفها. تعتمد عملية تمييز المصطلحات المحتملة الأولية في نظام TermRaider على المعالجة اللغوية المسبقة (تجزئة الجمل، تصنيف أقسام الكلام، إزالة الزوائد والعودة إلى أصل الكلمة، وتجزئة العبارات الاسمية)، التي يجري تنفيذها عادة في منصة GATE بواسطة أداة ANNIE أو أداة TwitIE (رغم إمكانية استخدام أدوات أخرى بدلاً من ذلك). بعدها تُستخلص المصطلحات المحتملة من النص بواسطة القواعد النحوية التي تفرض بدورها قيودًا على العبارات الاسمية، مثل استثناء بعض الكلمات المستبعدة المتكررة. أخيرًا، يُطبق نموذج tf/idf على المكنز، وهو ما يعطينا درجة تدل على مدى أهمية كل مصطلح محتمل في كل مستند. بعد ذلك يجري اختيار جميع المصطلحات المحتملة الحاصلة على درجة tf/idf أعلى من قيمة الحد الأدنى التي سبق تحديدها يدويًّا (تحدد هذه القيمة لتكون معامل وقت التشغيل) كمصطلحات.

إضافة إلى ذلك، يُطبق شكلان رئيسان إضافيان من أشكال نموذج tf/idf داخل نظام TermRaider. تضم قيمة tf/idf المعززة معلومات عن الكلمات المندرجة (hyponyms) تحت المصطلحات. المبدأ المعتمد هنا هو أن المصطلحات التي تندرج تحتها كلمات أخرى يُرجح أن تكون مصطلحات صحيحة. تمثل الدرجة الحد الأقصى لقيمة tf/idf المعززة المحلية الخاصة بالمصطلح المحتمل، وتحتسب هذه القيمة عن طريق الجمع بين درجة tf/idf الخاصة بالمصطلح المحتمل وبين درجات tf/idf الخاصة بجميع الكلمات المندرجة (hyponyms) تحت المصطلح المحتمل التي يُعثر عليها حول تلك الحالة (occurrence). هناك شكل آخر وهو درجة كيوتو (Kyoto) لأهمية النطاق [148]، التي تضم أيضًا عدد الكلمات المندرجة المتمايزة لكل مصطلح محتمل يرِد في المكنز بأكمله. مرة أخرى، يستند ذلك إلى المبدأ الذي ينص على أن المصطلحات التي توجد كلمات مندرجة تحتها هي مصطلحات صحيحة على الأرجح.

تستخدم طريقة NC-value [149] منهجية مشابهة، وتُستخدم كأساس لأدوات من قبيل TerMine[1]. هذه الطريقة مبنية على أساس نموذج tf/idf في المصطلحات المحتملة التي تُستخرج بطريقة مشابهة لأداة TermRaider، لكن جرى تطويرها عبر إضافة معلومات تتعلق بتكرار التوارد المشترك (co-occurance) مع الكلمات السياقية. بدورها تضيف منهجية TRUCKS [150] خصائص إضافية عن طريق تمييز الأجزاء المهمة في النص المحيط بالمصطلح، وقياس مدى قوة ارتباطها بالمصطلحات المحتملة ذات الصلة.

٦-٣-٢ المنهجيات التي تستخدم المعرفة السياقية

تأخذ المنهجيات التي تستخدم المعرفة السياقية في الاعتبار الكلمات الموجودة في سياق المصطلحات المحتملة من أجل المساعدة في تصنيفها. يمكن استخدام أنواع مختلفة من المعرفة، إما بصورة فردية أو بصورة مجتمعة. في بعض الأحيان تُستخدم هذه المعلومات من أجل استثناء مصطلحات معينة من كونها مصطلحات محتملة. لكنها تُستخدم في غالبية الحالات على شكل أوزان تساعد في تصنيف المصطلحات.

تتعلق المعرفة المصطلحية بحالة الكلمات السياقية. الكلمة السياقية التي تكون أيضًا مصطلحًا من المرجح أن تكون مؤشرًا أفضل يدل على كونها مصطلحًا مقارنة بكلمة سياقية ليست مصطلحًا. يعتمد هذا الأمر على الفكرة القائلة: إن المصطلحات تميل لأن تظهر مجتمعة في النص. على سبيل المثال، في منهجية TRUCKS [150] يجري توليد وزن لكل مصطلح محتمل بناءً على التكرار الإجمالي للمصطلح مع المصطلحات الأخرى الموجودة في سياقه.

تعتمد المعرفة النحوية على الكلمات الحدودية (boundary words)، أي الكلمات التي تسبق المصطلح المحتمل أو تليه مباشرة. تشترط منهجية كلمة الحاجز (barrier word approach) [151, 152] أخذ المصطلح بعين الاعتبار فقط عند وجود فئات نحوية معينة تسبق المصطلح المحتمل أو تليه. هناك أنظمة أخرى تخصص وزنًا لكل فئة نحوية من الكلمات السياقية المباشرة بناءً على تحليل تكرار التوارد. على سبيل المثل،

١- للتجميع، - للفصل، * و + و ؟ للتكرار.

يكون الفعل الذي يرد مباشرة قبل مصطلح محتمل مؤشرًا أفضل بكثير من الناحية الإحصائية على مصطلح حقيقي مقارنة بالنعت. بعد ذلك يُعطى كل مصطلح محتمل وزنًا نحويًّا يُحتسب عن طريق جمع أوزان الفئات لجميع الكلمات السياقية الحدودية الواردة معها.

تعتمد المعرفة الدلالية على فكرة تضمين المعلومات الدلالية المتعلقة بالسياق. يعتمد ذلك على مبدأ ينص على أن الكلمات الموجودة في السياق التي تحمل وجه شبه كبير بالمصطلح المحتمل من المرجح أن تكون مهمة أو ذات صلة. يمكن حساب التشابه بعدة طرق. راجع القسم ٦-٤ لقراءة بعض الأمثلة.

٦-٤ استخراج العلاقات

بعد استخلاص المصطلحات ذات الصلة، يجب توليد العلاقات الموجودة بينها. في الآونة الأخيرة، اقتُرحت العديد من منهجيات استخراج العلاقات، وتركز هذه المنهجيات على مهمة تطوير الأنطولوجيات (التعلم والتمديد والتعبئة). تهدف هذه المنهجيات إلى تعلم العلاقات التصنيفية القائمة بين المفاهيم، بدلاً من العناصر المعجمية. يختلف نوع استخراج العلاقة المطلوب لتطوير الأنطولوجيات قليلاً عن مهمة استخراج العلاقات التي تناولناها في الفصل الرابع، حيث كان التركيز في تلك المهمة على العلاقات غير التصنيفية، مثل مؤلفي الكتب، بينما نحن مهتمون هنا بالعلاقات التصنيفية من قبيل الكلمات المندرجة (hyponymy) (مثال: التفاح أحد أنواع الفاكهة).

٦-٤-١ أساليب التجميع

تهدف أساليب التجميع إلى تنظيم المصطلحات وفق تسلسل هرمي يمكن تحويله مباشرة إلى أنطولوجيا، وذلك باستخدام أسلوب من أساليب قياس المسافة بهدف إنشاء مجموعة من المصطلحات أو الدمج بينها. يقيس هذا الأسلوب مدى شبه مصطلح معين بمصطلح آخر أو بمجموعة مصطلحات أخرى، على سبيل المثال، يمكن استخدامه لحساب الحالات (instances) الأكثر نموذجية لمفهوم معين، مثل المفهوم الأقرب إلى الحالة المركزية (الحالة «المتوسطة» الافتراضية في المجموعة). هذه

المنهجية تتطلب أولاً اختيار قياس مسافة دلالي وخوارزمية تجميع مناسبين. المرجع [153] يحتوي على استعراض جيد للمنهجيات المختلفة ويمكن الرجوع إليه. تشمل أمثلة أساليب التجميع حيز المتجهات (vector space) [154]، والشبكات الترابطية [155] ومنهجيات المجموعات النظرية [156].

٦-٤-٢ العلاقات الدلالية

تقوم العلاقات الدلالية المبنية على الأنطولوجيا على مفهوم ينص على أن الكلمات المترابطة دلاليًّا ترد أو تظهر على مقربة بعضها من بعض داخل الأنطولوجيا مقارنة بالكلمات التي يكون ترابطها أضعف. قد يكون هذا الأمر مفيدًا في عملية وضع المصطلحات داخل الأنطولوجيا بصورة صحيحة وفي مهام إزالة غموض المصطلحات. هناك عدد من المقاييس المختلفة المستخدمة لقياس درجة الترابط، ويمكن تصنيفها إلى ثلاثة أنواع رئيسة، وهي: الأساليب المبنية على التكرار، والأساليب المبنية على القواميس، والأساليب المبنية على الأمثلة. يمكن الاطلاع على وصف أطول لهذه الأساليب في [154]. نورد هنا تلخيصًا لبعض من أبرزها.

تُستخدم الأساليب المبنية على التكرار بكثرة في عمليات استرجاع المعلومات، وهي مبنية على الخصائص الإحصائية للكلمات الموجودة في المكانز. تضم هذه الأساليب قياس جاكارد (Jaccard) الموزون [158] وأساليب التوارد المشترك البسيط (مثال: تكرار التوارد المشترك والمعلومات المتبادلة ونسبة الترابط) والأساليب القائمة على المتجهات، وهذه الأساليب تقيس درجة التشابه بين الكلمات باستخدام حاصل الضرب النقطي أو الجداء القياسي (product dot) أو دالة جيب التمام (cosine function) أو المسافة الإقليدية (Euclidean distance) بين متجهين يمثلان سياقات الكلمات المقدمة في تعريفهما. يجري حساب المتجه الخاص بالسياق عن طريق إضافة متجهات معلومات التوارد المشترك الخاصة بالكلمات الموجودة في التعريف، ويمكن إيجاد ذلك عن طريق توارد بسيط.

تعتمد الأساليب المبنية على القواميس على قاموس أو أنطولوجيا مهيكلة وفق تسلسل هرمي، حيث تُحدد أوزان العُقد الموجودة في التسلسل بشكل عام بناءً على

التكرار أو الاحتمالية. تشمل الأساليب الشائعة لحساب أوجه التشابه المسافة المفاهيمية والمسافة الدلالية والأشكال المختلفة. المسافة المفهومية [159] هي مسافة الممر الأقصر الرابط بين الحالات (instances) كلها في التسلسل الهرمي. تُقاس المسافة الدلالية [160] بواسطة محتوى المعلومات الخاص بـ Abstraction Most Specific Common (MSCA) -الفئة الأكثر تحديدًا في التسلسل الهرمي التي تندرج تحتها كلتا الفئتان. يُحتسب محتوى المعلومات من خلال تقدير احتمال ورود الفئة داخل أحد المكانز. يمكن كذلك أخذ عمق العقدة في التسلسل الهرمي بعين الاعتبار، وذلك لأن العُقد التي توجد في مستويات عميقة من التسلسل الهرمي تميل لأن تكون متشابهة بصورة كبرى.

تُستخدم الأساليب المبنية على الأمثلة بكثرة في الترجمة الآلية، وتهدف إلى اختيار التجربة الأكثر شبهًا بمشكلة معينة. تجمع هذه الأساليب عادة بين هياكل ذات تسلسل هرمي ومجموعة من الأمثلة المأخوذة من أحد المكانز. تشمل هذه الأساليب رسوم الخصائص الموزونة [161] وتقارب الكلمات [162] وخوارزميات التطابق الأفضل [163] والمسافة الدلالية الموزونة المعتمدة على الأمثلة [164].

تُستخدم الأساليب الدلالية المستندة إلى المكانز في الغالب في مهمة استخراج العلاقات بهدف إنشاء الأنطولوجيات. تقوم هذه الأساليب على فكرة أن الكلمات المترابطة دلاليًّا ترد معًا في النص. علاوة على ذلك، تتوارد مثل هذه الكلمات بتكرار أكبر مقارنة بالكلمات غير المترابطة (أو التي يكون ترابطها أقل قوة). على سبيل المثال، التفاح أكثر ارتباطًا بالبرتقال من الأحذية، وذلك لأن كليهما من أنواع الفاكهة بينما الأحذية ليست كذلك. لذا فإننا نتوقع أن ترد كلمة تفاح في النص نفسه بتكرار أكبر مع كلمة برتقال مقارنة بكلمة أحذية. عن طريق مقارنة تكرارات هذين التواردين، يمكننا تحديد أن التفاح والبرتقال بينهما ارتباط أقوى من الارتباط الموجود بين التفاح والأحذية. تتميز المنهجيات المستندة إلى المكانز بكونها قائمة بذاتها ولا تتطلب أي مصادر خارجية، وهذا يعني أنها مناسبة للغاية للنطاقات المتخصصة، وتميل نحو ضمان أن تكون المعلومات مناسبة لذلك النطاق. غير أن استخدام المعلومات الناتجة عن مكنز كهذا قد تؤدي إلى حدوث انحراف إحصائي، وقد يكون هناك فجوات في تغطية المكنز. يبين الجدول رقم ٦-١ بعضًا من إيجابيات وسلبيات المنهجية القائمة على المكانز [157].

الجدول ٦-١ : سلبيات وإيجابيات المنهجية المستندة إلى المكانز لاستخراج العلاقات الدلالية

الإيجابيات	السلبيات
استخدام أمثلة حقيقية على اللغة	قد تكون الأساليب غير موثوقة
معلومات مصممة خصيصًا للنطاق	قد تكون التغطية غير كافية
المعلومات الإحصائية متوفرة	الحاجة إلى مكنز ضخم
	وجود فجوات في التغطية
	قد تكون المعلومات غامضة

٦-٤-٣ الأنماط المعجمية النحوية

أنماط هيرست هي مجموعة من الأنماط المعجمية النحوية التي تشير إلى وجود علاقات شمول (hyponymic relations) [165]، وقد استُخدمت هذه الأنماط على نطاق واسع لإيجاد العلاقات بين المصطلحات وإنشاء الأنطولوجيات. تُستخدم الأنماط أيضًا في كل من برنامجي Text2Onto وSPRAT (انظر أدناه). في العادة تحقق مستوى عاليًا من الدقة، إلا أن الاسترجاع متدنٍّ جدًّا لديها، وبعبارة أخرى تتميز بالدقة الشديدة لكنها لا تغطي سوى مجموعة فرعية فقط من الأنماط الممكنة لإيجاد الكلمات الشاملة (hypernyms) والكلمات المشمولة (hybonyms). ولهذا السبب فإنها عادة ما تُجمع مع أنواع أخرى من الأنماط.

يمكن وصف أنماط هيرست (Hearst patterns) بواسطة القواعد التالية، حيث تعني NP عبارة اسمية بينما تحمل التعبيرات القياسية معانيها المعتادة(١):

1. such NP as (NP,)* (or|and) NP

مثال: works by such authors as Herrick, Goldsmith, and Shakespeare.....

2. NP (,NP)* (,)? (or|and) (other|another) NP

مثال: Bruises, wounds, or other injuries.........

3. NP (,)? (including|especially) (NP,)* (or|and) NP

مثال: All common-law countries, including Canada and England......

1- http://www.bbc.co.uk/news/technology-27711109

هناك حالات لا تعمل فيها هذه الأمثلة. على سبيل المثال، يمكن للمرء استخراج كلمة الإيطاليين ككلمة مشمولة (hybonym) في عبارة أوروبيون الواردة في جملة الأوروبيون، لاسيما الإيطاليين، لكن ينبغي على المرء عدم استخراج الديمقراطيين ككلمة مشمولة (hybonym) في عبارة الرؤساء الأمريكيون الواردة في جملة الرؤساء الأمريكيون، ولا سيما الديمقراطيين.

وبناء على ما سبق، قام بيرلاند وتشارنياك [166] أيضًا بتطوير بعض الأنماط للتعامل مع أسماء الأجزاء (meronymy)، على سبيل المثال، لاستخراج أن عداد السرعة هو أحد أجزاء السيارة. فيما يلي اثنان من أمثلة الأنماط:

1. NN's NN
   … building's basement…

2. NN of DET (JJ|NN)* NN
   … basement of a building…

كما أن نظام SPRAT الذي جرى تطويره كأحد ملحقات منصة GATE والذي سيرد شرحه في القسم ٦-٦ يشمل أيضًا أنماطًا إضافية.

٦-٤-٤ الأساليب الإحصائية

في حين تنتج الأنماط المعجمية النحوية في العادة علاقات نموذجية (مثل الشمول (hyponymy)) بين المصطلحات، يمكن إيجاد علاقات تركيبية أو نَسَقية (مثل المتلازمات اللفظية (collocations)) باستخدام أساليب إحصائية. يعدُّ أسلوب المعلومات المتبادلة النقطية [167] من الأساليب المشهورة التي تقيس الاعتماد المتبادل بين اثنين من المتغيرات. يستخدم هذا الأسلوب عادة في لغويات المكانز كدالة أهمية لحساب المتلازمات اللفظية (Pointwise Mutual Information) [168]. لإيجاد العلاقات، يمكننا استخدام هذا الأسلوب لقياس مدى قوة الارتباط بين اثنين من المصطلحات داخل المستند نفسه أو المكنز [169].

٦-٥ إثراء الأنطولوجيات

في العادة لا تكون الأنطولوجيات ثابتة بل دائمة التطور. في البداية، قد تجري إضافة مفاهيم (أنواع) جديدة أو حذفها أو تحريكها. عند إجراء مثل هذه التغييرات، ينبغي أن تنعكس أيضًا على الحالات (instances) والعلاقات (الخصائص). ثانيًا، قد يتعين إضافة حالات جديدة أو حذفها أو تحريكها لكي تصبح الأنطولوجيا أكثر كمالاً أو لتصحيح المشكلات الموجودة. لإدخال تغييرات هيكلية على الأنطولوجيا، ينبغي إعداد آليات مبدئية للتعامل مع هذا الأمر، وذلك للحيلولة دون فقدان معلومات صحيحة (مثال: تحريك الحالة إلى مستوى أعلى في التسلسل الهرمي عند حذف المفهوم الذي تنتمي إليه تلك الحالة). غير أن هذه التغييرات لا تتطلب في العادة تكنولوجيا معالجة اللغات الطبيعية. لهذا السبب سوف نحصر النقاش هنا في الأساليب المستخدمة لإثراء الأنطولوجيات عبر إضافة حالات وعلاقات جديدة.

من بين الأسباب الرئيسة التي تجعل الأنطولوجيا غير مكتملة في العادة وجود مشكلة البيانات المتناثرة. عند إنشاء أنطولوجيا باستخدام أحد المكانز، لن تكون المعلومات التي يحتوي عليها المكنز كاملة أبدًا -ولذلك لا نتوقع احتواء أي مجموعة من النصوص على جميع المصطلحات الموجودة في نطاق معين أو أن تُظهر أنماطًا معجمية نحوية لجمع العلاقات بين المصطلحات. يوجد هذا النوع من اختناق اكتساب المفردات (lexical acquisition bottleneck) بكثرة في مهام معالجة اللغة، وغالبًا ما تُحل هذه المشكلة باستخدام أساليب التجميع. لغرض إثراء الأنطولوجيات، يمكن استخدام الأطر الدلالية. تعود هذه الفكرة إلى أواخر الستينات مع ظهور الفرضية التوزيعية [167] التي طرحها هاريس (أي أن الكلمات التي تظهر في السياق نفسه تميل لأن تحمل معانِيَ متشابهة)، والأعمال التي تمت في السبعينات [170, 171] التي ركزت على تحديد مجموعات من أنواع الكلمات الخاصة باللغات الفرعية باستخدام أنماط نحوية مستقاة من نصوص خاصة بالنطاق. على وجه الخصوص، ظلت الأبحاث في هذا المجال تُستخدم في نطاقات محددة كالطب، حيث عادة ما يوجد عدد صغير نسبيًّا من الهياكل النحوية في تقارير المرضى مثلاً. تكون الهياكل هنا بسيطة للغاية، وتكون الجمل قصيرة وغير غامضة نسبيًّا: وهو ما يجعل عملية المطابقة بين الأنماط النحوية أسهل بكثير. تتمثل

الفكرة الأساسية في أنه يمكن إنشاء أنواع كلمات دلالية (مجموعات) عن طريق معاينة مجموعات من العناصر المعجمية التي توجد في بيئات نحوية محددة. على سبيل المثال، قام (هيرشمان وآخرون) [172] بتطوير نوع (type) جديد في مجال التقارير السريرية هو العلامة أو العرض، يتكون من عناصر معجمية مثل نزلة برد خفيفة، حمى، سعال طفيف، الخ، وذلك بواسطة جمع حالات العناصر المعجمية التي توجد كمفعولين بهم للفعل أصيب، بالإضافة إلى الفاعل مريض. يظهر في الجدول رقم 2-6 مثال على ما أطلقوا عليه صيغة المعلومات (information format).

منذ ذلك الوقت، أجريت الكثير من الأعمال حول اكتساب المعرفة الدلالية وفقًا لمنهجية مشابهة. على سبيل المثال، قام روشا [173] بدور ريادي في استخدام أطر الحالات لما يسميه نماذج تعريف الأحداث (تشبه إلى حد بعيد الأطر المستخدمة في عملية استخراج المعلومات لتعريف الأحداث، كما تُستخدم في تقييمات مؤتمرات تقييم الرسائل). من بين الأمثلة على أطر الحالات هذه المثال الظاهر في الجدول رقم 3-6.

الجدول ٦-٢: صيغة المعلومات الخاصة بالنوع (العلامة) أو (العرض)

الفاعل	الفعل	المفعول به
المريض	أصيب	نزلة برد خفيفة
المريض	أصيب	حمى
المريض	أصيب	سعال طفيف
المريض	أصيب	صداع

الجدول رقم ٦-٣: مثال لإطار الحالة الذي طرحه روشا

الفتحة	الحشوة
العملية:	أشعة سينية للصدر
الرابط:	يظهر

٦-٦ أدوات تطوير الأنطولوجيات

في هذا القسم سوف نشرح عددًا من الأدوات المستخدمة عادة لإنشاء الأنطولوجيات وإثرائها آليًّا اعتمادا على أساليب معالجة اللغات الطبيعية.

٦-٦-١ TEXT2ONTO

أداة TEXT2ONTO [174] من أولى الأدوات وأشهرها لتطوير الأنطولوجيات آليًّا. تقوم هذه الأداة باستخراج المترادفات على أساس الأنماط، وتجمع بين منهجيتي التعلم الآلي ومهام المعالجة اللغوية الأساسية مثل تجزئة الجمل وإزالة الزوائد والعودة إلى أصل الكلمة والتحليل النحوي السطحي. ونظرًا لكونها مبنية على إطار منصة GATE، فإنها توفر مرونة من حيث خيارات الخوارزميات التي يمكن تطبيقها.

٦-٦-٢ SPRAT

نظام SPRAT (أداة لتمييز الأنماط الدلالية وإضافة الشروحات إليها) [175]. يعد نظام SPRAT مثالا لأنظمة تطوير الأنطولوجيات لنطاق الأسماك، على الرغم من إمكانية تطبيق منهجيته في النطاقات الأخرى. هذا النظام قادر على إنشاء أنطولوجيا جديدة من الصفر، أو تعديل أنطولوجيا موجودة مسبقًا، وهو مبني على مبدأ الأنماط المعجمية النحوية. مقارنة بنظام Text2Onto، يضم هذا النظام عددًا أكثر من الأنماط المعجمية النحوية، لكنه لا يستخدم التجميع والتحليل النحوي الإحصائي لاستخراج العلاقات. هذا يعني أن النظام يصدر كمية أقل من البيانات، لكن يحتمل أن يكون أكثر دقة.

٦-٦-٣ FRED

نظام FRED هو أداة إلكترونية لتحويل النصوص إلى أنطولوجيات مترابطة جاهزة للبيانات، وذلك باستخدام التحليل النحوي. يجمع النظام بين نظرية تمثيل الخطاب (DRT) ودلالات الإطار اللغوي وأنماط تصميم الأنطولوجيات (ODP). هذا النظام مبني على أساس أداة Boxer [177] اللغوية التي تقوم بتوليد التمثيلات الدلالية الرسمية للنص، بناءً على دلالات الأحداث. وفي حين تركز الأدوات الأخرى في العادة بصورة رئيسة على مساعدة المستخدم في التعرف على المصطلحات الأساسية التي ينبغي

إضافتها إلى الأنطولوجيا، يختلف نظام FRED في كونه يهدف إلى تقديم أنطولوجيات وبيانات مترابطة جاهزة للاستخدام.

٦-٦-٤ الإنشاء شبه الآلي للأنطولوجيات

في مجال هندسة الأنطولوجيات، ظهرت أنماط تصميم الأنطولوجيات [178] كطريقة لمساعدة مطوري الأنطولوجيات في نمذجة أنطولوجيات OWL وفقًا لأسلوب من الأعلى إلى الأسفل. أنماط تصميم الأنطولوجيات (ODPs) هي في الأساس مجموعات من الأنماط المفاهيمية المصممة لمساعدة المستخدمين في تصميم أو تنقيح الأنطولوجيات. جرى أيضًا تطوير أدوات لدعم إعادة الاستخدام شبه الآلي لهذه الأدوات [179]. تستخدم هذه الأدوات نصوصًا ذات صلة بالنطاق كمدخلات لها، بينما تكون مخرجاتها مجموعة من أنماط تصميم الأنطولوجيات لحل احتياجات الأنطولوجيات الأولية. تجري المقابلة بين أنماط تصميم الأنطولوجيات وصياغات اللغات الطبيعية من خلال الأنماط المعجمية النحوية.

ركزنا في هذا الفصل حتى الآن على وصف أنماط إنشاء الأنطولوجيات من المكانز وفقًا لأسلوب من الأعلى إلى الأسفل. من البدائل المتاحة للمستخدمين ممن ليسوا من الخبراء عند إنشاء أنماط تصميم الأنطولوجيات (ODPs) هي استخدام تراكيب الجمل أو اللغات المقيدة (restricted languages) المصممة خصيصًا لجعل الأنطولوجيات أكثر قابلية للقراءة والفهم من قبل الآخرين. تشمل الأمثلة على ذلك لغةAttempto Controlled English (ACE) [180] ولغة Rabbit [181] ولغة Sydney OWL Syntax ولغة CLOnE [182] (لغة تعديل الأنطولوجيات المقيّدة) [183]. يبين الجدول رقم ٦-٤ عددًا من أمثلة الجمل الموجودة في هذه اللغات. تتمثل الفكرة الرئيسة التي تقوم عليها هذه اللغات المقيّدة في السماح للأفراد ممن ليسوا من الخبراء بالتعبير عن احتياجاتهم الخاصة بنمذجة الأنطولوجيات وفقًا لمجموعة معينة من القواعد النحوية. على المرء أن يكون على دراية مسبقة بالمصطلحات والعلاقات التي يرغب في نمذجتها، حيث تكمن المشكلة في تحويل هذه المصطلحات والعلاقات إلى الشكل الأنطولوجي الصحيح. على سبيل المثال، عند استخدام لغة CLoNE، بإمكان الخبير في النطاق استخدام واجهة لغة طبيعية لتحويل النص الموجود لديه إلى أنطولوجيا

بسيطة -مع كتابة النص في واجهة المستخدم، يجري تحويله بشكل آلي (باستخدام عملية معالجة اللغات الطبيعية) إلى أنواع وعلاقات في الأنطولوجيا. غير أن الصعوبة تكمن في أن على المستخدم كتابة النص وفقًا لأسلوب محدد جدًّا، وذلك حسب اللغة المقيّدة المستخدمة.

الجدول ٦-٤: أمثلة على اللغات المقيّدة المستخدمة في إنشاء الأنطولوجيات

اللغة	أمثلة الجمل
ACE	Every river-stretch has-part at-most 2 confluences.
Rabbit	Every Bourne is a kind of stream.
Sydney Syntax	The classes petrol station and gas states are equivalent.
CLoNE	Projects have string names

٦-٧ خاتمة

ناقشنا في هذا الفصل مهمة إنشاء الأنطولوجيا آليا مع عرض مكوناتها الرئيسة، وهي التعلم والتعبئة والتنقيح. وفي حين يوجد الكثير من المنهجيات المتبعة لإنشاء الأنطولوجيات آليًّا، إلا أننا ركزنا هنا على الأساليب المستندة إلى تقنيات معالجة اللغات الطبيعية والتي تُبنى على ما ناقشناه من مكونات تتألف منها معالجة اللغات الطبيعية التي سبق أن شرحناها في الفصول السابقة، وهي المعالجة المسبقة وتمييز كيانات الأسماء واستخراج العلاقات. كما ركزنا هنا بصفة خاصة على استخراج المصطلحات نظرًا لأنها المكون الأساسي في عملية إنشاء الأنطولوجيا، وكذلك على الأساليب المستخدمة في ترتيب هذه المصطلحات وفق تسلسل هرمي. يعدُّ استخراج العلاقات مكونًا رئيسا آخر، ونظرًا لأننا قد سبق أن شرحنا هذا المكون بشكل مفصل في القسم 4-6، فقد اقتصرنا هنا على عرض تلخيص لأهم أنواع العلاقات التي تعد مفيدة لعملية توليد الأنطولوجيا، كما سلطنا الضوء على الأنماط المعجمية النحوية. وفي الختام أشرنا إلى العديد من العناصر المترابطة في عملية إنشاء الأنطولوجيا، ومنها إنشاء الأنطولوجيا شبه الآلي، كما قدمنا بعض الأمثلة على الأدوات المستخدمة عادة في هذا المجال.

الفصل السابع
تحليل المشاعر

٧-١ مقدمة

من أهم جوانب فهم النص تمييز وتصنيف الآراء والمشاعر والعواطف. قد تتفاوت هذه المهمة بين تصنيف تقييمات المستخدمين لمنتجات معينة (هل أعجب هذا المنتج المستخدم أم لا؟ ما خصائص المنتج التي أعجبته/ لم تعجبه؟) وفهم المشاعر والعواطف التي تحملها التغريدات، وتتبع الآراء مع مرور الوقت وتمييز آراء المؤثرين والقادة وإعداد الخلاصات بناءً على الآراء. يشرح هذا الفصل المكونات الأساسية لأدوات تحليل المشاعر النموذجية، كما يقدم تشكيلة متنوعة من شتى الأساليب التي يمكن استخدامها، ويعطي أمثلة للتطبيقات الموجودة في الواقع العملي في مختلف المجالات، ويبرز كيف يمكن إدراج مهمة تحليل المشاعر ضمن تطبيقات أشمل تستخدم لتحليل محتوى شبكات التواصل الاجتماعي.

تحليل المشاعر (داخل النص) هي عملية تتعلق بتحليل النص من أجل فهم آراء الناس. نحن لسنا هنا بصدد تحليل المشاعر داخل الأشكال الأخرى للوسائط كالصور والفيدوهات، وذلك لكونها لا تندرج تحت أساليب معالجة اللغات الطبيعية. في أبسط الحالات، يعني ذلك فهم ما إذا كان أحد الأشخاص يتحدث بأسلوب إيجابي أو سلبي عن شيء ما، لكن بالطبع يمكن أن تأخذ الآراء طابعًا أكثر غموضًا، فقد تعبر عن مختلف أنواع العواطف وقد تختلف تلك العواطف في شدتها (هل الشخص معجب بشيء ما قليلاً أو كثيرًا، هل هو خائف، مصدوم، غاضب، مرتاح، متفاجئ على نحو إيجابي الخ؟). يمكن أن تعبر العواطف أيضًا عن الشعور تجاه جوانب محددة في منتجات أو حدث ما، الأمر الذي يؤدي بصفة عامة إلى وجود قدر من التناقض (كأن تكون معجبًا ببعض العناصر وغير معجب ببعضها الآخر).

قد تكون أدوات تحليل المشاعر مفيدة للغاية في كل القطاعات الصناعية تقريبًا. من الأمثلة النموذجية على ذلك تقييمات المنتجات، فقد يبحث شخص يرغب في شراء كاميرا عن التعليقات والتقييمات الموجودة على شبكة الإنترنت، بينما قد يرغب شخص آخر قام بشراء كاميرا بوضع تعليق على المنتج والحديث عن تجربته؛ بينما يمكن لمصنعي الكاميرات الحصول على ملحوظات من عملائهم، وهو ما قد يساعدهم في تطوير منتجاتهم أو خدماتهم و/ أو تعديل استراتيجيتهم التسويقية. إن محاولة تحليل هذه

التقييمات والآراء يدويًّا غالبًا ما تكون غير مجدية، ولا سيّما بالنسبة للشركات الكبرى التي قد تصلها ملايين التقييمات الخاصة بكل منتج. في حين يوجد في المواقع الرسمية لتقييمات المنتجات أنظمة لحساب التقييمات بواسطة النجوم، إلا أن المعلومات الأهم من حيث الفائدة للمستخدم غالبًا ما توجد في النص الحر، ما يعني أن تجميع الدرجات العددية ليس كافيًا لرؤية الصورة الكاملة. أضف إلى ذلك أن التعليقات التي تُنشر على شبكات التواصل الاجتماعي كتويتر غالبًا ما تتطلب استجابة فورية، ومع ضرورة عدم الاعتماد على الأنظمة الآلية بالكامل للتجاوب مع تلك التعليقات، إلا أن أدوات تعدين الآراء قد تساعد في الإبلاغ عن المشكلات الخطيرة، أو إبراز الاتجاهات الجديدة. قد تستفيد أنظمة الإجابات على الأسئلة أيضًا إلى حد بعيد من مكونات تعدين الآراء، وذلك من أجل التعامل مع أسئلة من قبيل «ما أفضل مطعم ياباني في لندن؟» أو ما شابه. قد يحاول المرء أيضًا الرد على الأسئلة التي تتطلب فهمًا أكثر تعقيدًا، كسؤال يقول: «ما الكاميرا الفضلى من حيث عمر البطارية؟»

وفي حين يمكن أن تكون تقييمات وآراء العملاء أهدافًا واضحة لأدوات تعدين الآراء، وبالنظر لتركيز جزء كبير من الأبحاث عليها (يعود ذلك جزئيًّا إلى وجود حاجة واضحة، لكنه أيضًا بسبب سهولة إنشاء أطقم خاصة بالتدريب والاختبار مكونة من كميات ضخمة من البيانات باستخدام أنظمة التقييم كمعيار ذهبي)، إلا أن هناك العديد من الاستخدامات الأخرى لأدوات تعدين الآراء. من بين المهام المُهمة الأخرى أمور مثل فهم المشاعر السياسية والاجتماعية تجاه الحكومات والأحداث والانتخابات وما إلى ذلك. تقليديًّا، كانت تُجرى هذه التحليلات بواسطة استطلاعات الرأي (مثل YouGov في المملكة المتحدة)، غير أنها باهظة الثمن وتستهلك الكثير من الوقت. يشكل التحليل التنبئي أو التوقعي (predictive analysis) على وجه الخصوص سوقًا ضخمًا، بداية بمعرفة الأفلام التي ستفوز بجوائز الأوسكار وغيرها من الجوائز (وهو ما يؤدي بالتالي إلى زيادة الإيرادات)، مرورًا بالتحقيق في كيفية تأثير المزاج العام على سوق الأسهم وعمل التوقعات بناءً على الأحاديث الدائرة على شبكات التواصل الاجتماعي. يمكن استخدام التحليلات الاجتماعية أيضًا لشرح الاختلافات المهمة، ليس عبر الارتباطات الصريحة (الأشخاص الذين يحبون السفر قد يرغبون في شراء منتجات السفر) فحسب، بل أيضًا من خلال الارتباطات الضمنية غير الصريحة (على سبيل

المثال: الأشخاص الذين يقومون بشراء منتجات نايك يميلون أيضًا لشراء منتجات أبل).

تقوم أدوات تعدين الآراء بأخذ قطعة من النص كمُدخلات، وتعطي مخرجاتٍ على شكل معلومات تحدد ما إذا كان النص يتضمن آراء، وما طبيعة الآراء التي يعبر عنها (إيجابية، سلبية، ...الخ)، ومدى قوة الرأي، وأيضًا احتمال وجود معلومات أخرى مثل الموضوع الذي يتعلق به الرأي، ومن صاحب الرأي، وتعطي نوعًا من أنواع تلخيص الآراء بعدة جمل أو تعبيرات. سنناقش هذه المهام الفرعية بمزيد من التفصيل في الفقرة ٧-٣.

قد تبدو مهمة تعدين الآراء بسيطة للوهلة الأولى، فقد يبحث نظام بسيط وغير معقد عن وجود كلمات إيجابية وسلبية (مثل أكره، جيد، سيئ-..الخ) ومن ثمّ يقوم بتوليد الرأي الناتج وفقًا لذلك. في الممارسة العملية، تكون مهمة تعدين الآراء أكثر تعقيدًا من ذلك، حتى في حال مهام كشف قطبية الرأي (polarity detection) (معرفة ما إذا كانت عبارة ما إيجابية أو سلبية). يعود السبب في ذلك كما سبق أن رأينا في هذا الكتاب إلى كون اللغات الطبيعية شديدة التعقيد والغموض. ينطبق هذا الأمر على وجه التحديد على شبكات التواصل الاجتماعي، حيث تتركز مهمة تعدين الآراء. يلجأ الناس إلى استخدام مصطلحات غير معتادة في شبكات التواصل الاجتماعي لوصف مشاعرهم، ويقومون بإضافة تعبيرات سلبية إلى ما يكتبونه من تعبيرات، ولا يستخدمون قواعد النحو والإملاء على النحو الصائب، ويستخدمون العبارات الشرطية وعبارات المشاعر كأسئلة، وقد يكونون ساخرين أو متهكمين، وقد يفترضون أن القارئ يملك معرفة إضافية بالعالم المحيط به تمكنه من فك شفرة المعنى من دون إعطاء إشارات واضحة (على سبيل المثال: تكون الإشارات إلى فولديمورت (Voldemort) أو هتلر (Hitler) سلبية بشكل عام). هذا يعني في الغالب ضرورة إجراء تحليل لغوي معقد لفك رموز المعنى بصورة صحيحة، كما سنرى في القسم ٧-٢ والقسم ٧-٣.

أخيرًا، علينا أن نوضح في هذا القسم نقطة تتعلق بالمصطلحات. من الناحية النظرية، الآراء والمشاعر أمران مختلفان، ومن ثم فهناك اختلاف بين تعدين الآراء وتحليل المشاعر تبعًا لذلك. تعبر المشاعر عادة عن درجة قطبية معينة (إيجابي، سلبي،

أو محايد). على سبيل المثال، عبارة «أظن أن فستانك جميل» تحمل مشاعر إيجابية أعبر عنها. قد تعبر الآراء عن شيء ما أكثر شمولاً، على سبيل المثال، عبارة «أظن أنها ستمطر غدًا» هي رأي أعبر عنه أنا بشأن الطقس، لكنها لا تعبر عن مشاعر محددة إيجابية كانت أو سلبية. غير أن «الرأي» يمكن أن يُستخدم أيضًا ليعني مشاعر إيجابية أو سلبية، وفي المثال الأول، أعبر عن رأي إيجابي يتعلق بفستانك.

في المراحل المبكرة لبحوث تعدين الآراء، استُخدم مصطلح «تعدين الآراء» ليعني شيئًا أكثر شمولاً بكثير مما هو عليه الآن، في حين كان تحليل المشاعر يُستخدم للإشارة تحديدًا إلى مهمة كشف قطبية الرأي. غير أنه خلال السنوات الأخيرة بات المصطلحان كلاهما يُستخدمان بشكل تبادلي، وبالأخص في الحالات التي تم فيها إنشاء مهام فرعية ومهام جانبية (على سبيل المثال: كشف ما إذا كان شيء ما يحمل رأيًا أو لا، وكشف وجود المشاعر وإلى أي مدى يمكن الوثوق بالآراء، وما إلى ذلك -راجع الأقسام التالية). في هذا الفصل، نستخدم تعبير «تعدين الآراء» ليشمل مهام تتضمن كشف ما إذا كان شيء ما يعبر عن مشاعر معينة، وما هي درجة القطبية في تلك المشاعر، وما مدى قوتها، ومن صاحب الرأي، وبماذا يتعلق الرأي، وما طبيعة العواطف التي يجري التعبير عنها. نحن لا نسعى لتصنيف الآراء كتعبيرات خالية من الحقائق وذات مشاعر محايدة (كما هو الحال في مثال الطقس) والتمييز بينها وبين تعبيرات الحقائق (مثال: «إنها تُمطر»).

٧-٢ المشكلات الموجودة في تعدين الآراء

قد تستخدم منهجية مبسطة لتحليل المشاعر معجمًا يضم كلمات تحمل آراء (جيد، سعيد، حزين، ...الخ) وتجميع هذه الكلمات من النص قيد التحليل (كجُملة أو تغريدة أو مستند) من أجل اتخاذ قرار بشأن درجة القطبية العامة. في حقيقة الأمر، تستخدم العديد من المنهجيات الأساسية هذا الأسلوب بالذات، وتحصل على درجات مقبولة. لكن حتى لو أخذنا بعين الاعتبار مشكلات من قبيل النفي («جيد» مقابل «غير جيد»)، تبقى هناك العديد من الفروق الدقيقة التي تعيق هذا النوع من التحليل المبسط. على سبيل المثال، قد تغير الجمل الشرطية المعنى تغييرًا كبيرًا («إن خسرت أسكتلندا المباراة، فإنها ستكون كارثة»). قد يختلف الرأي أيضًا وبشكل كبير من حالة إلى أخرى، وذلك

تبعًا لصاحب الرأي والموضوع الذي يتعلق به. تحمل عبارة «إن خسارة أسكتلندا للمباراة أمر رائع» ضمنيًّا مشاعر إيجابية يعبر عنها كاتبها بشأن نتيجة المباراة، لكنها تحمل أيضًا نوعًا من المشاعر السلبية تجاه أسكتلندا. على الجانب الآخر، نحن لا نتوقع أن تكون أسكتلندا أو المشجعون الأسكتلنديون سعداء بهذه النتيجة. حتى الكلمات البذيئة والمصطلحات السلبية يمكن أن تستخدم استخدامًا إيجابيًّا، إن توفر السياق الصحيح، فالبريطانيون بالتحديد غالبًا ما يشيرون إلى أصدقائهم مستخدمين مصطلحات في غاية السلبية دون أن يكونوا سلبيين تجاههم بأي شكل من الأشكال (على سبيل المثال: نعت شخص ما بكلمة mucker يعدُّ نوعًا من التحبب، لكن هذه الكلمة تعني حرفيًّا الشخص الذي يقوم بإزالة النفايات).

على المرء أيضًا أن يكون حذرًا بخصوص التمييز بين رأي بشأن شخص أو شيء ما، وبين حدث يتعلق بذلك الشخص أو الشيء. على سبيل المثال، التعبير عن الحزن أو الصدمة لوفاة شخص ما ليس مؤشرًا على كراهية ذلك الشخص، حتى على الرغم من كون مضمون الرسالة سلبيًّا بصفة عامة، غير أن العديد من أدوات تحليل المشاعر تخطئ هنا لكونها لا تميز بين الأمرين.

قد تكون هناك أيضًا صعوبة في التعامل مع السخرية، لكن المحتوى الساخر يغلب على محتوى شبكات التواصل الاجتماعي. في البداية، يجب على النظام التعرف على السخرية عند وجودها، وهي مهمة لا تكون سهلة دائمًا، حتى بالنسبة لشخص يملك معرفة سياقية كبرى. ثانيًا، يجب على النظام فهم كيفية تأثير السخرية أو التهكم على درجة قطبية الرأي، فقد تقوم بعكس درجة القطبية المتوقعة للعبارة أو الجملة بأسرها، أو لجزء صغير منها فقط، أو حتى عدة جمل [184]. وفي حين قد تبدو القدرة على كشف السخرية هدفًا ثانويًّا، إلا أن الآثار المترتبة عليها مهمة للغاية، ففي عام 2014م، أعلنت المخابرات الأمريكية عن وجود خطط لديها لشراء برمجيات للمراقبة الآنية لمستخدمي شبكات التواصل الاجتماعي، وهي خطط تتضمن تحديدًا القدرةَ على كشف السخرية[1].

1- http://www.emotion-research.net

٧-٣ مهام تعدين الآراء الفرعية

يتضح من النقاش السابق أن هناك عددًا من المشكلات في مهام تعدين الآراء ينبغي معالجتها من قبل الأدوات التي تقوم بهذه المهمة آليًّا. يمكن تقسيم هذه المهام إلى مجموعة من المهام الفرعية الاختيارية التي يمكن للأدوات استخدامها. سنعطي فيما يلي وصفًا موجزًا لهذه الأدوات والأساليب التي يمكن استخدامها.

٧-٣-١ كشف القطبية

كشف القطبية (polarity detection) هي مهمة تتعلق بتحديد ما إذا كانت عبارة ما إيجابية أو سلبية أو محايدة. في بعض الأحيان، تكون هذه المهة جزءًا من مهمة كشف الآراء (هل تحمل هذه العبارة رأيًا؟)، حيث يشير الحياد إلى أن العبارة لا تحمل رأيًا، بينما يشير التصنيفان الآخران إلى أن العبارة تحمل رأيًا. تقوم الأنظمة الأخرى أولاً بتصنيف العبارات إلى مهام فرعية. يمكن تقييم هذه المهام أيضًا كمهمة واحدة أو مهمتين منفصلتين. تقوم الأنظمة الأخرى أولاً بتصنيف العبارات إلى عبارات تحمل آراء وعبارات لا تحمل أي آراء، ومن ثم تقوم بتصنيف العبارات التي تحمل آراء مرة أخرى في مهمة فرعية منفصلة. يمكن تنفيذها كمهمة واحدة أو كمهمتين منفصلتين. بعض الأنظمة تميز بين الحياد وعدم وجود مشاعر، وغالبًا ما يكون الأمر كذلك عند استخدام النظام في المستندات الطويلة. تكون هذه المستندات عادة محايدة بسبب وجود عدد متساوٍ من العناصر الإيجابية والسلبية. من الأمثلة على ذلك موقع تقييمات يوجد فيه تقييم بدرجة 5 /3 نجوم، حيث يمكن اعتبار هذه الدرجة إيجابية وسلبية بصورة متساوية، وذلك لوجود بعض النقاط الجيدة والسيئة المتعلقة بالمنتج. بدلاً من ذلك، تُستخدم المشاعر المحايدة في بعض الأحيان لوصف الحالات التي يعبر فيها الكاتب بوضوح عن بعض المشاعر، لكن لا يتضح فيها ما طبيعة المشاعر تحديدًا. في تلك الحالات، يختلف عدم وجود مشاعر عن حياد المشاعر. غير أن الأدوات اليدوية والآلية المستخدمة لإضافة التعليقات والشروحات تجد صعوبات كبيرة في التمييز بين الحالتين، ولا سيّما في المستندات القصيرة، ولذا يتم الجمع بين الحالتين دون أي تمييز.

٧-٣-٢ كشف هدف الرأي

غالبًا ما تكون معرفة كون الرأي إيجابيًّا أو سلبيًّا أمرًا غير كافٍ، ما لم نعرف أيضًا بالتحديد الموضوع الذي يكون الرأي إيجابيًّا أو سلبيًّا بشأنه. كما ناقشنا سابقًا، محبة شخص ما تختلف اختلافًا كبيرًا عن محبة موته. وبالمثل قد يكون الإعجاب بسمة من سمات شخص أو شيء ما (شعر الشخص، لون سيارته، ...الخ) مختلفًا كثيرًا عن الإعجاب بالشخص أو الشيء ككل. تتعلق مهمة كشف الهدف (target detection) بتمييز الأمر الذي يتعلق به الرأي، وتتبع منهجيتين رئيستين في هذا الصدد. تعمل المنهجية الأولى وفق مفهوم من الأعلى إلى الأسفل (top-down) وتُستخدم عندما يكون الهدف محددًا سلفًا وعادة ما يكون الهدف سمة أو خاصية من خصائص شيء ما توجد في إحدى الأنطولوجيات أو غيرها من أنظمة التصنيف (على سبيل المثال: الفنادق لديها خصائص مثل الغرف وخدمة الطعام والموقع؛ والكاميرات لديها سعر وحجم وعمر بطارية وما إلى ذلك). سنورد شرح تعدين الآراء المستند إلى الخصائص بواسطة الأنطولوجيات في القسم 6-7. المنهجية الثانية هي منهجية تتم وفق مفهوم من الأسفل إلى الأعلى (bottom-up)، حيث تكون الأهداف المحتملة غير معروفة سلفًا، لكنها تؤخذ من النص بشكل آلي. في العادة تتألف هذه المنهجيات من مصطلحات أو كيانات أو أحداث سبق تحديدها في مرحلة سابقة من مراحل عملية معالجة اللغات الطبيعية. لكن تظل مهمة ربط الرأي بالكيان الصحيح تحديًا يتطلب مزيدا من الدراسات حوله، ومجرد استخدام المنهجيات المستندة إلى المسافات غير كافٍ إلى حد بعيد، والأنسب اتباع منهجية بدوافع لغوية من أجل الحصول على أفضل النتائج (أي استخدام التحليل النحوي أو على الأقل تجزئة النص لضمان الحفاظ على العلاقة الصحيحة بين الكلمات التي تحمل آراء والهدف المطلوب). لكن تبقى هذه المهمة غير سهلة، ويعود سبب ذلك جزئيًّا إلى الأخطاء التي تقع في مهمة التحليل النحوي (ولا سيّما في نصوص شبكات التواصل الاجتماعي)، وجزئيًّا بسبب تعقيد التركيبات. توجد أمثلة على المنهجيات المستندة إلى الكيانات في [185] وفي [186]. كما توجد أمثلة على المنهجيات ذات الأهداف المحددة سلفًا، والتي تُعرف أيضًا باسم كشف المواقف (stance detection)، في [187] وفي[188].

٧-٣-٣ كشف صاحب الرأي

مثلما هو الحال مع مهمة كشف هدف الرأي، تتعلق مهمة كشف صاحب الرأي (opinion holder detection) بالتعرف على الشخص الذي يحمل الرأي المشار إليه. قد يكون الأمر بسيطًا في العديد من الحالات، على سبيل المثال في آراء العملاء التي عادة ما يكون صاحب الرأي هو الشخص الذي يكتب التقييم، على الرغم من أن الأمر قد لا يكون بالبساطة نفسها في حالات أخرى («الكتاب أعجب صديقي، لكنني أجده مملاً للغاية»). في الحالات التي لا يكون كاتب النص صاحب الرأي، يكون الأمر متعلقًا بحالات الكلام المنقول (يستخدم على نحو فضفاض للإشارة إلى أفعال من قبيل التفكير، الشعور-...الخ). يمكن التعرف على هذه الأنواع من التراكيب باستخدام تحليل لغوي ذي جودة عالية قادر على التعرف على أسماء أو أنواع أصحاب الآراء المحتملين (عادة ما يكونون أشخاصًا أو مؤسسات) والتصنيفات الدلالية للأفعال (تفكير، شعور، قول، ...الخ) والأنماط الدلالية لنموذج مثل صاحب-رأي-فعل-رأي (opinion – opinion_verb–holder). الحالة الأخرى هي المثال المبين أعلاه («الكتاب أعجب صديقي») حيث يتعين تمييز فاعل الفعل الذي يحمل الرأي وتصنيفه على أنه صاحب الرأي. في التغريدات، قد يكون صاحب الرأي أيضًا كاتب تغريدة أصلية جرت إعادة تغريدها. هنا، ينبغي الحذر في تحديد ما إذا كان المرء يرغب في التعرف على الكاتب الأصلي للتغريدة أو الشخص الذي قام بإعادة نشرها، أو كليهما، وتصنيفه على أساس أنه صاحب الرأي. لاحظ أن الأخير أمر مثير للجدل إلى حد ما، ولا سيّما عندما يرغب المرء في إبراز عبارة مثيرة للجدل. وكما هو الحال مع كشف هدف الرأي، تعدُّ مهمة كشف الكيان (entity detection) خطوة أولى مهمة في عملية تمييز الكاتب، على الرغم من أنه قد يكون من الضروري تحديد العبارات الاسمية المتعلقة بالأشخاص والمؤسسات، مثل «صديقي».

٧-٣-٤ تجميع المشاعر

يمكن تحديد المشاعر بعدة مستويات، وعادة ما يكون ذلك على مستوى الجملة/ العبارة أو على مستوى المستند/ المشاركة. عادة ما تتكون التغريدات من جملة واحدة ومن ثم يجري التعامل معها على أنها تندرج تحت الفئة الأولى، لكنها في بعض الأحيان

تتكون من عدة جمل. وبالتالي، يجري التعرف على الرأي عادة على مستوى التغريدة، لكن باستخدام منهجيات على مستوى الجملة، وذلك بسبب تعبير كل تغريدة عن رأي واحد في العادة. في الغالب تبدأ عملية تحليل المشاعر التي تطبق على المقالات أو المشاركات الأطول (مثل تقييمات الأفلام) بتعدين الآراء على مستوى الجملة، والعمل على أساس جملة أو تعبير واحد وتقسيم التقييم أو المقال إلى عدد من الآراء المختلفة على الأرجح حول الخصائص المختلفة لهدف الرأي (على سبيل المثال: «كان الطعام شهيًّا، لكن الخدمة كانت بطيئة للغاية»). يأتي نقاش مفهوم تعدين الآراء وفقًا للخصائص بمزيد من التفصيل في القسم ٧-٦، ويلاحظ هذا المفهوم عادة في تحليل مواقع تقييم المنتجات.

هناك منهجيتان رئيستان لتجميع المشاعر. تتمثل المنهجية الأولى، وهي الأكثر شيوعًا، في الجمع بين جميع الدرجات الإيجابية والسلبية لكل جملة أو عبارة، وتقديم درجة موحدة إجمالية، وهو ما يؤمل أن يتوافق مع التصنيف النجمي، إن وجد. في الواقع، تُستخدم التصنيفات النجمية كبيانات تدريبية لمثل هذه الأنظمة، على الرغم من أن ذلك قد يطرح إشكالية بسبب عدم كون الدرجة الموحدة الإجمالية والتصنيف النجمي متوافقين دائمًا (قد يقوم المرء بإعطاء تقييم ذي ٤ نجوم، ومن ثم استخدام النص الحر فقط لشرح النقاط السلبية). بالنسبة للمستندات مثل المقالات والمدونات، أو مجموعات التعليقات، ليست هناك دائمًا علاقة مباشرة بين النقاط الإيجابية والسلبية المجمّعة. تقول بعض النظريات: إن المشاعر المحايدة تحمل في الواقع قيمة إيجابية تزيد قليلاً عن الحالات التي تكون فيها المشاعر غائبة، ولذا تجري موازنة هذه الأمور بهذه الطريقة. وعلى نحو مماثل، تميل المشاعر السلبية عادة للتفوق على المشاعر الإيجابية (يميل الناس لنشر آرائهم عندما لا يكونون سعداء بشأن أمر ما). هناك طريقة ثانية أقل شيوعًا للحصول على درجة موحدة للمشاعر عندما يتعلق الأمر بالمستندات الطويلة، وهي طريقة الجمع بمرور الوقت (collect-as-you-go)، حيث يجري البحث داخل المستند كلمة بكلمة وتحديث الدرجة تبعًا لذلك. يُعرف هذا الأسلوب بالتحليل الجماعي (بدلاً من التحليل التجميعي) [185].

٧-٣-٥ المكونات اللغوية الفرعية الإضافية

لمعالجة بعض المشكلات المتبقية التي ورد ذكرها سابقًا، قد تستفيد أدوات تعدين الآراء من عدد من المكونات اللغوية الفرعية الإضافية. التحليل النحوي، أو على الأقل تجزئة النص، هما مكونان مفيدان في تجزئة الجمل إلى أجزاء صغيرة، وذلك من أجل إيجاد العلاقات الصحيحة بين المكونات مثل الآراء والأهداف وأصحاب الآراء. الأسلوب الأبسط للقيام بذلك هو تجزئة الوحدات وفقًا لعلامات الترقيم وكلمات التنسيق، على الرغم من أنها عملية ليست محمية ضد الفشل بأي حال من الأحوال. يعطي التحليل النحوي نتائج أفضل لأنه يتيح استخراج علاقات التبعية الصحيحة (راجع الفصل الثاني)، لكنه غالبًا ما يطرح إشكالية من حيث الأداء في نصوص شبكات التواصل الاجتماعي وبالأخص التغريدات، وذلك بسبب غياب الاستخدام الصحيح للقواعد النحوية في النص.

من المفيد القدرة على التعرف على الهياكل اللغوية كالأسئلة والعبارات الشرطية، وذلك لأنها قد تؤثِّر في النص الذي يتضمن رأيًا إلى حد بعيد. وفي حين قد تحمل الأسئلة مشاعر (ضمنية في العادة)، إلا أن هذا الأمر غير معتاد إلى حد ما. عند طرح سؤال «هل تعتقد أن هذا الفستان جميل؟»، فهذا لا يعني في العادة وجود مشاعر إيجابية أو سلبية لدى السائل. وبالمثل، يعبر السؤالان «لو كان هذا الفستان أزرق اللون لكان جميلاً» و»لو كنت أرغب في الحصول على فستان رخيص، لكنت اشتريت فستانًا مختلفًا» كلاهما يعبر عن مشاعر مركّبة، لذا ينبغي إيلاء عناية خاصة هنا. في الواقع، بإمكان المرء أن يذهب أبعد من ذلك ويتعرف على قواعد محددة تتعلق بالمشاعر بناءً على نوع العبارة الشرطية: تطبق هذه القواعد، على سبيل المثال، في أنظمة GATE من أجل تحليل المشاعر]190[، حيث تكون عملية إضافة مثل هذه المكونات الإضافية سهلة للغاية.

تشكل العبارات البذيئة حالة خاصة؛ ولذلك يجب أن نوليها اهتماما خاصا لأن بعضها قد يبدو أنه تعبير سلبي ولكن ليس الأمر كذلك في سياق الكلام. تندرج العبارات البذيئة في العادة في معاجم المشاعر السلبية، لكن الناس لا يستخدمون العبارات البذيئة بطريقة سلبية دائمًا. في حقيقة الأمر، تُستخدم هذه العبارات عمومًا كنوع من أدوات تعزيز المشاعر، ولا سيّما عندما ترد في النص كمُعَدِّلات (modifiers)

لصفات أو أسماء إيجابية أو سلبية –على سبيل المثال: «bloody awful» (سيء جدا) مقابل «bloody good» (جميل جدا) –.

كما ذكرنا سابقًا، يُعد كشف السخرية مجالا آخرَ من المجالات التي ينبغي إيلاؤها عناية خاصة. من الناحية التقليدية، كانت أنظمة الآراء تتجاهل السخرية والتهكم نظرًا لصعوبة التعرف عليهما بصورة آلية، إلا أنهما كانا في الآونة الأخيرة موضوع بحوث متزايدة [184 ,191]. تشمل المنهجيات المستخدمة عادة تدريب أدوات التصنيف على التغريدات التي تضم علامات تصنيف (هاشتاغ) من قبيل –سخرية و–ساخر، والتغريدات التي لا تشمل مثل هذه العلامات [192].جرى تحقيق نجاح معتبر مع مثل هذه الأساليب من حيث التعرف على ما إذا كانت التغريدة ساخرة أم لا، لكنَّ قدرًا يسيرًا من الأبحاث تناول المشكلة المتعلقة بكيفية تأثير السخرية على القطبية نفسها، وذلك لأن هذا الأمر ليس بسيطًا (راجع [184] للاطلاع على نقاش حول هذه المشكلة).

٧-٤ كشف العواطف

أدوات تعدين الآراء المستخدمة للمهام العملية تبتعد على نحو متزايد عن الأدوات العادية المستخدمة لكشف المشاعر الإيجابية/ السلبية وتسير نحو اتباع منهجية قائمة على العواطف، حيث تصنّف هذه المنهجية النصوص التي تحمل الآراء وفقًا للعواطف التي تعبر عنها، ويمكن الاطلاع على مثال لذلك في [193]. يعود السبب الرئيس في ذلك إلى أنه يعدُّ الخيار الأجدى للأغراض العملية. على سبيل المثال، تفضل الشركات عمومًا أن تعرف بالتحديد ما إذا كان الناس يشعرون بالخوف أو الغضب تجاه منتج معين، بدلاً من مجرد شعورهم بمشاعر سلبية تجاهها. هناك مسار بحثي آخر تناول التلازم بين العواطف (ولا سيّما الخوف) والتغييرات في أسعار أسواق الأسهم [194]. قد تحتوي العواطف على قِيَم دقيقة (fine-grained) تُعبّر على شكل مفاهيم من أنطولوجيات ذات تعريف جيد.

غير أن مهمة تحديد مجموعة كاملة وواضحة من العواطف هي مهمة صعبة. جرت عدة محاولات لتحديد عدد من المعايير (راجع، على سبيل المثال [195] و //:http

www.emotion-research.net)، لكن لا يوجد حتى الآن إجماع على مجموعة أساسية من العواطف. من التمثيلات التي يشيع اقتباسها عجلة بلوتشيك للعواطف المبينة في الشكل رقم ٧-١. تعد هذه العجلة محاولة لإظهار كيفية ارتباط العواطف المختلف بعضها ببعض، لكن ربما تبدو معقدة جدًّا لدرجة تجعلها غير مناسبة لتمثيل عملية تمييز العواطف. تُظهر العجلة ثمانية عواطف أساسية ثنائية القطب كما هو مبين في الدائرة التي تأتي في المرتبة الثانية من حيث العمق، وهي الفرح مقابل الحزن، والغضب مقابل الخوف، والثقة مقابل الاشمئزاز، والمفاجأة مقابل الترقب. تتمثل الفكرة التالية بعد ذلك في أنه مثلما هو الحال مع الألوان، يمكن التعبير عن المشاعر الأساسية بدرجات متفاوتة في شدتها، كما يمكن المزج بينها لتشكيل عواطف أخرى. على سبيل المثال، المزج بين الترقب والفرح يعطيك التفاؤل، ونقيض ذلك هو الاستنكار. تعدُّ المشاعر الموجودة على طرفي النقيض مصدر القلق الأكبر، حيث يتوقع المرء أن يكون التشاؤم نقيض التفاؤل على سبيل المثال. وبالمثل، تصنف العجلة الغضب على أنه نقيض مفهوم الخوف الأساسي، كما تصنف الثقة كنقيض للاشمئزاز. حتى لو أخذنا هذه الفئات كنقطة بداية من دون الأخذ بعين الاعتبار التفاعل فيما بينها، هناك عدد من الفئات المتوقعة التي لا توجد في العجلة، إلا أن العواطف الأساسية الثماني تستخدم بكثرة لأغراض التصنيف الآلي.

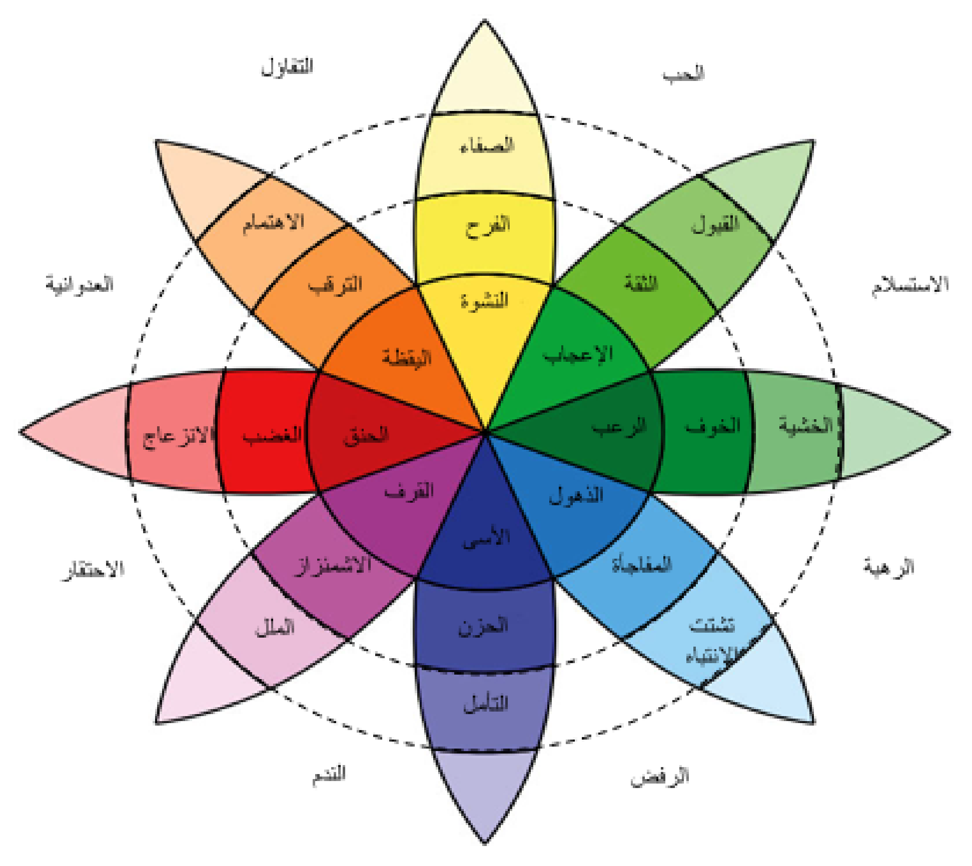
تستخدم قائمة مشاعر باروت المهيكلة على شكل شجرة [196] التي شرحها [197] للمرة الأولى، الفئات الأساسية التي طرحها بلوتشيك، لكنها تزيد عددًا بصورة مختلفة. تستخدم القائمة ثلاثة مستويات، ويظهر أول مستويين في الجدول رقم ٧-١.

هناك تمثيل آخر يسمى EARL (لغة تمثيل شروحات العواطف)، وقد جرى تطويرها خصيصًا لإضافة الشروحات والتعليقات إلى العواطف من قبل شبكة التفاعل بين الإنسان والآلة حول العواطف (HUMAINE[1])، وتصنف ٤٨ نوعًا من أنواع العواطف، كما هو مبيّن في الجدول رقم ٧-٢ والجدول رقم ٧-٣.

من النقاط المهمة التي ينبغي وضعها بعين الاعتبار هو أنه وخلافًا لقطبيات الآراء العمومية (إيجابي/ سلبي)، لا تكون نقائض العواطف بالضرورة سلبيات العواطف.

1- http://linguistic-lod.org/

على سبيل المثال، على الرغم من أن السعادة والحزن يعتبران في العادة شعورين متناقضين، وفي حين يمكن عمومًا إعادة صياغة العبارة «أنا لست سعيدًا» لتصبح «أنا حزين»، لكن على الجانب الآخر لا تحمل عبارة «أنا لست حزينًا» المعنى نفسه الذي تحمله عبارة «أنا سعيد». في حقيقة الأمر، يمكن تعميم هذا المفهوم بشكل أكبر: نفي العواطف الإيجابية يكون سلبيًّا في العادة، لكن نفي العواطف السلبية قد يكون محايدًا في كثير من الأحيان بدلاً من أن يكون إيجابيًّا. هذا يعني أن الأسلوب المعتاد المتمثل في قلب أو عكس القطبية عند مصادفة التعابير السلبية ليس بالضرورة حلاًّ جيدًا عندما يتعلق الأمر بتمييز العواطف. وكما يبدو، فإن هذا الأمر لم يتم تناوله ضمن أدبيات البحث.



الشكل رقم ٧-١: عجلة بلوتشيك للعواطف (الرسم لـMachine Elf ١٧٣٥. مرخص بموجب الملكية العامة المشاعة).

الجدول ٧-١ : تصنيف باروت للعواطف

العواطف الرئيسة	العواطف الثانوية
الحب	الحنان الشهوة/ الرغبة الجنسية التوق
الفرح	المرح التلذذ الرضا الفخر التفاؤل الافتتان ارتياح
المفاجأة	**المفاجأة**
الغضب	الانفعال السخط الحنق الكراهية الاشمئزاز الحسد عذاب
الحزن	المعاناة الكآبة خيبة الأمل العار الإهمال تعاطف
الخوف	الرعب التوتر

٧-٥ أساليب تعدين الآراء

مع كون تعدين الآراء ميدانًا جديدًا من ميادين البحث، إلا أن الكثير من الأبحاث قد جرت خلال العقد الماضي (وما بعده) حول أساليب تحديد الآراء وتصنيفها. توجد مراجعة شاملة ومفصّلة للأساليب التقليدية لكشف المشاعر آليًّا في [198]، ومنها العديد من المكونات الفرعية. بصفة عامة، يمكن تقسيم تلك الأساليب إلى أساليب مبنية على المعاجم وأساليب مبنية على التعلم الآلي. تعتمد الأساليب المبنية على المعاجم على معجم مشاعر، وهو عبارة عن مجموعة من مصطلحات المشاعر المعروفة والمجمّعة سلفًا. تستخدم منهيجات التعلم الآلي الخصائص النحوية و/أو اللغوية، فيما يشيع كثيرًا استخدام منهجيات هجينة، حيث تلعب معاجم المشاعر دورًا مهمًّا في غالبية هذه الأساليب. حتى الأساليب البسيطة يمكن أن تكون فعالة للغاية، ومن الأمثلة على ذلك تحديد قطبية تقييمات المنتجات عبر تحديد قطبية النعوت التي تظهر فيها (تفيد التقارير أن هذه المنهجية حققت دقة أكبر من أساليب التعلم الآلي المحض بنسبة ١٠٪ [199]). لكن مثل هذه الأساليب الناجحة نسبيًّا غالبًا ما تفشل عند نقلها إلى نطاقات أو أنواع نصوص جديدة، وذلك بسبب كونها غير مرنة بخصوص غموض مصطلحات المشاعر. يمكن أن يتغير المعنى الذي يحمله السياق الذي يُستخدم فيه المصطلح، ولا سيّما النعوت الموجودة في معاجم المشاعر [200]. على سبيل المثال، تعدُّ السيارة الهادئة من الممتلكات الإيجابية، لكن الأمر ليس كذلك عمومًا بالنسبة لمنبه هادئ. إضافة إلى ذلك، برهنت عدة تقييمات مدى أهمية المعلومات السياقية [201]، [202]، وحددت الكلمات السياقية ذات التأثير الأعلى على قطبية المصطلحات الغامضة [203]. هناك صعوبة أخرى تتمثل في عملية إنشاء قواميس المشاعر المستهلكة للوقت، على الرغم من طرح عدد من الحلول مثل أساليب التعهيد الجماعي (crowdsourcing).

الجدول ٧-٢: تمثيل EARL للعواطف السلبية

قوي	الغضب الانزعاج الازدراء الاشمئزاز التهيج	لامبالي	الملل اليأس خيبة الأمل الجرح الحزن	أفكار سلبية	شك الحسد الإحباط الشعور بالذنب عار
فقدان السيطرة	الهم الإحراج الخوف العجز الضعف القلق	الهياج	الإجهاد صدمة التوتر		

الجدول ٧-٣: تمثيل EARL للعواطف الإيجابية

حيوي	لتسلية لبهجة الانتشاء الإثارة السعادة الفرح لمتعة	مهتم	المودة التعاطف الصداقة الحب	أفكارإيجابية	الشجاعة الأمل لفخر الرضا لثقة
		هادئ إيجابي	لسكون الاطمئنان الاسترخاء الارتياح لصفاء	تفاعلي	الاهتمام الكياسة المفاجأة

في الآونة الأخيرة، بدأت أساليب تعدين الآراء تركز على شبكات التواصل الاجتماعي، إلى جانب بروز توجه جديد نحو تطبيق هذه الأساليب على نحو استباقي بدلاً من تطبيقها كآليات تأتي كرد فعل. قد تكون لفهم طبيعة الرأي العام بهذه الطريقة آثار على توقع الأحداث المستقبلية بالنسبة للحكومات ووسائل الإعلام الراغبة في

معرفة ردود الأفعال التي ستحدث نتيجة للأحداث والسياسات، وكذلك بالنسبة للأشخاص الراغبين في توقع أداء أسواق الأسهم وأمور أخرى كثيرة. غير أن تكييف هذه الأدوات لتتعامل مع شبكات التواصل الاجتماعي بعيدٌ كل البعد عن أن يكون مهمة يسيرة في أغلب الأحيان، كما سنشرح في الفصل الثامن. على وجه الخصوص، لا تعمل مكونات المعالجة اللغوية المسبقة في وسائل التواصل الاجتماعي في الغالب بشكل جيد، بالإضافة إلى أن الرسائل القصيرة المتبادلة على تويتر تفتقر إلى معلومات سياقية مفيدة، كما يوجد فيها العديد من الأخطاء الإملائية، وهو ما يعني أن كلمات المشاعر معرضة للفقدان. علاوة على ذلك، يشيع استخدام اللغة العامية وغالبًا ما تكون الرسائل غامضة (يكون ذلك مقصودًا في بعض الأحيان).

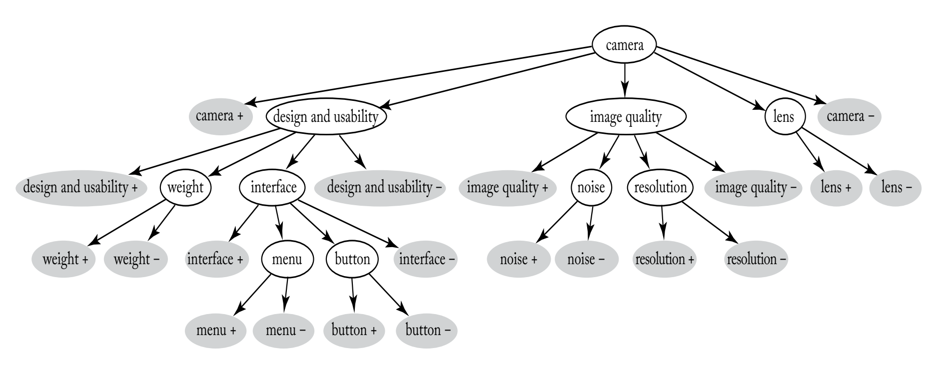
تستخدم الغالبية العظمى من أساليب تعدين الآراء أسلوب التعلم الآلي، ويعود ذلك جزئيًّا إلى سرعة إعداده وسهولته، وأيضًا بسبب النتائج المعقولة التي يمكن الحصول عليها بأقل قدر من الجهد. تكون المنهجيات الخاضعة للإشراف مفيدة بصفة خاصة عندما تتوفر كميات ضخمة من بيانات التدريب، مثل آراء المستخدمين التي تجمع بين نظام تقييم صريح ونص حر. غير أن مثل هذه المنهجيات لا تتكيف بصورة جيدة مع التغريدات وغيرها من أشكال محتوى شبكات التواصل الاجتماعي [204]، ولا سيما المحتوى الذي يكون خاصًّا بنطاق معين. في الحالات الخاصة، يمكن إنشاء بيانات التدريب باستخدام علامات التصنيف (الهاشتاغ) أو رموز الانفعالات (emoticons)، لكنها غالبًا ما تشكل جزءًا صغيرًا من البيانات ذات الصلة؛ نظرًا لأن معظم الأشخاص لا يستخدمون هذه الرموز في تغريداتهم. لهذا السبب، ركز قسم من الأبحاث على تكييف أساليب التعلم الآلي مع النطاقات الجديدة [205]، لكن هذه الأبحاث تركز في العادة على استخدام كلمات مفتاحية (keywords) مختلفة مع أنواع نصوص متشابهة، على سبيل المثال، تقييمات المنتجات المتعلقة بالكتب مقابل تقييمات الأجهزة الإلكترونية. عندما يتعلق الأمر بمهام تعدين الآراء الهادفة، خصوصًا في التطبيقات الصناعية بدلاً من الأبحاث التخمينية، يُفضل عادة استخدام قاعدة معرفة لأنها تتيح للمطورين تخصيص أداة تعدين الآراء لتتلاءم مع المهمة، ومن الأمثلة على ذلك التركيز بشكل خاص على أهداف وأنواع الآراء، بدلاً من مجرد السعي للعثور على تغريدات أو تصنيفات عواطف إيجابية وسلبية ذات طابع عام.

تستخدم منهجيات تعدين الآراء المستندة إلى المعرفة في العادة مهام المعالجة اللغوية المسبقة، كما سبق شرحه في الفصل الثاني، بالإضافة إلى قواميس جغرافية (معاجم كيانات أسماء gazetteers) تضم معاجم المشاعر، بالإضافة إلى بعض القواعد التي تحكم طريقة الجمع بين درجات المشاعر (sentiment scores) وغيرها من الخصائص اللغوية (كالارتباط بالكيانات لغرض تمييز الأهداف، وتعديل الدرجات عند العثور على كلمات سلبية أو ضمائر أو ما شابه، وكذلك التبعيات السياقية وما إلى ذلك). وبالتالي يكون تعديل هذه الأساليب شديد السهولة على المستخدم عند العثور على أخطاء، على سبيل المثال في حال اكتشاف عدم وجود كلمات أو عبارات مشاعر داخل المعجم، أو عند استخدام كلمات المشاعر بطريقة معينة، أو عند استخدام تعبيرات لغوية معينة، وما إلى ذلك. توجد أمثلة على أدوات تعدين الآراء المعتمدة على المعرفة المستخدمة في العادة في أدوات مثل GATE، و[206] VADERو[207] SO-CAL.

٧-٦ تعدين الآراء والأنطولوجيات

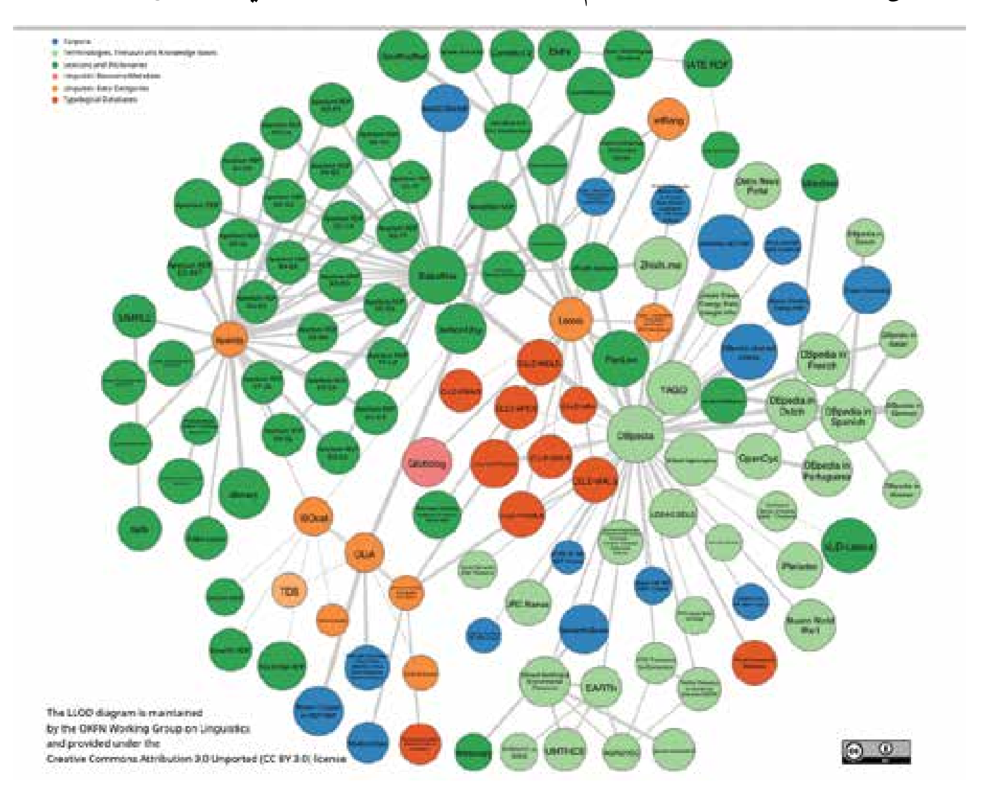
تحليل المشاعر على مستوى المفهوم مصطلحٌ يُستخدم عادة للإشارة إلى المنهجيات التي تتجاوز التحليل على مستوى الكلمات، وتركز بدلاً من ذلك على التحليل الدلالي بناءً على الأنطولوجيات أو البيانات المترابطة أو غيرها من المصادر الدلالية. نعني هنا بالتحليل الدلالي أن هذه المنهجيات تبتعد عن الاستخدام التقليدي والصريح للمعاجم ومعلومات التوارد المشترك (co-occurance) لتنتقل إلى منهجية جديدة تعتمد على الخصائص الضمنية المرتبطة بمفاهيم اللغات الطبيعية [208]. على سبيل المثال، تعد أداة SentiWordNet مصدرًا معتمدًا على أداة WordNet التي تضيف معلومات المشاعر (درجات خاصة بالإيجابية والسلبية والموضوعية) لكل مجموعة من مجموعات المترادفات (synset) في نظام WordNet. لهذا السبب، تتيح عملية ربط كلمات المشاعر التي يُعثر عليها في النص بنظام SentiWordNet العثور بسهولة على المترادفات والأشكال المختلفة للكلمات. جرى تصميم تحديات التحليل الدلالي على مستوى المفهوم (CLSA) في عامي 2014 و2015 بالذات لتشجيع تطوير تقنيات تعدين الآراء الدلالي، وتظهر عددًا من الأمثلة الممتازة [208، 209]، ومن المقرر استمرار سلسلة المؤتمرات حتى عام 2016.

يوجد مثال على هذه الأنظمة في [210]، ويقوم هذا النظام بنمذجة نطاق التقييمات الإلكترونية باستخدام أنطولوجيا معبأة بحالات (instances) مأخوذة من قاعدة المعرفة DBpedia. جرى توسيع الحالات (instances) المأخوذة من مجموعة البيانات المعجمية في قاعدة المعرفة DBpedia [211] باستخدام الأطر السياقية (contextual frames) (أي استخدام مجموع الكلمات المحيطة بمصطلح معين للعثور على مصطلحات جديدة ذات صلة كما سبق شرحه في الفصل السادس). معاجم المشاعر وثلاثيات المفاهيم (concept triples) المرتبطة بها مشمولة أيضًا (مثال: مشروب، بارد، إيجابية). تقوم الأنظمة الأخرى مثل]212[بترميز المصطلحات ذات الصلة بمفهوم معين (خصائص) داخل أنطولوجيا، وبعد ذلك تقوم بتوسيع نطاق مجموع المصطلحات عن طريق إضافة كلمات مرادفة وكلمات مندرجة (hyponyms) يُعثر عليها داخل النص. على سبيل المثال: -تكبير، عمر البطارية، تأخير غالق الكاميرا، الخ- هي مجموعة تضم الخصائص التي تشترك فيها جميع المنتجات التي تندرج تحت فئة كاميرا رقمية [213]. يُعرف هذا الأمر غالبًا باسم تعدين الآراء المعتمد على الخصائص. يعدُّ الشكل رقم 2-7 مثالاً آخر على أنطولوجيا مكونة من الخصائص في نطاق الكاميرات. نلاحظ أن غالبية هذه المنهجيات مصممٌ للتعامل مع النطاقات المغلقة كتقييمات المنتجات، حيث يمكن نمذجة المنتجات وخصائصها بسهولة. تزداد صعوبة استخدام هذا النوع من المنهجيات بصورة كبيرة عند تطبيقها على تعدين الآراء في النطاقات المفتوحة التي تكون فيها مجموعة أهداف الآراء الممكنة غير معروفة.



الشكل ٧-٢: قسم من أنطولوجيا خصائص الآراء، مقتبسة من عرض تقديمي بعنوان «استرجاع المعلومات: البحث عن الآراء في البرية»، -Opinion Retrieval: Looking for Opinions in the Wild- الدكتور جيورجوس بالتوغلو.

لكن من بين التحديات المعيقة لتطوير مثل هذه الأدوات الحاجة للدمج بين المصادر اللغوية الموجودة حاليًّا المستخدمة لتحليل المشاعر والمصادر الدلالية. تعدُّ مبادرة البيانات المفتوحة المترابطة اللغوية (Linguistic Linked Open Data Cloud (LLOD))[1] السحابية مثالًا على المبادرات التي تهدف إلى توفير موارد لغوية شبيهة بالبيانات المفتوحة المترابطة السحابية (Linked Open Data Cloud (LLOD))[2]، وذلك باستخدام مفردات من قبيل OWL وlemon وNIF للتعبير عنها، غير أن مهمة تحويل الموارد القديمة إلى هذا النظام ودمجها به ليست مهمة سهلة بأي حال من الأحوال.



الشكل ٧-٣: سحابة البيانات المفتوحة المترابطة اللغوية، اعتبارًا من شهر يناير ٢٠١٦ (جرى توليدها آليًّا من البيانات الموجودة في منصة Linghub وتقوم بصيانتها مجموعة العمل المعنية باللغويات التابعة لمؤسسة المعرفة المفتوحة (OKFN Working Group on Linguistics)).

1- http://lemon-model.net/

2- http://persistence.uni-leipzig.org/nlp2rdf/

٧-٧ أدوات تعدين الآراء

من بين أدوات تعدين الآراء الأكثر شعبية المستخدمة من قبل الباحثين وفي بعض التطبيقات الصناعية أداة Sentistrength [214] ويعود سبب ذلك بصورة رئيسة إلى كونها متاحة مجانًا وتعمل بصورة جيدة ويسهل إعدادها واستخدامها كأداة منفصلة أو ضمن تطبيقات أخرى. هذه الأداة مصممة لتقدير مدى قوة المشاعر الإيجابية والسلبية في النصوص القصيرة، وتتعامل بصورة جيدة مع اللغة غير الرسمية كالتي تستخدم في التغريدات. وخلافًا لمعظم الأدوات الأخرى، تقدم أداة SentiStrength اثنين من المؤشرات التي تدل على قوة المشاعر بصورة منفصلة، وهما مؤشر السلبية الذي يتراوح بين -١ و-٥ (حيث يدل -٥ على مؤشر سلبي للغاية)، ومؤشر الإيجابية الذي يتراوح بين ١ و٥ (حيث يدل ٥ على مؤشر إيجابي للغاية). تتوفر نسخة خاصة بنظام التشغيل ويندوز وكذلك نسخة مبنية بلغة جافا[1]، كما جرى دمجها في الآونة الأخيرة مع منصة GATE كملحق إضافي، مع العلم أن جميع النسخ قابلة للتخصيص عبر عدد من المعاملات (parameters) المختلفة. غير أن هذه الأداة تعاني من المشكلات المعتادة في أدوات تعدين الآراء الحالية، فهي تعمل جيدًا مع المشاعر الصريحة، لكنها لا تعمل بالجودة نفسها مع التعبيرات الأكثر تعقيدًا أو التي تتطلب قدرًا من المعرفة بطبيعة ما يجري في العالم، كما تعتمد الجودة إلى حد بعيد على جودة المعاجم الخاصة بها.

تحتوي معظم أطقم الأدوات الرئيسة الخاصة بمعالجة اللغات الطبيعية على مكونات خاصة بتعدين الآراء، أو يمكن على الأقل تطبيقها على هذه المهمة. تشمل هذه الأدوات NLTK وUIMA وLingpipe وطقم أدوات Stanford وGATE، وكذلك حزمة تعدين النصوص الخاصة بنظام R وأيضًا Weka وRapid Miner، وكلاهما لديه حزم خاصة بالتصنيف. تستخدم غالبية هذه الأنظمة أساليب التعلم الآلي (باستثناء منصة GATE التي تمتلك الاثنين) ولذا فهي تعتمد بشكل رئيس على جودة البيانات التدريبية والخصائص التي جرى اختيارها.

1- http://sentistrength.wlv.ac.uk/

٧-٨ خاتمة

في هذا الفصل، قدمنا شرحًا لمفهوم تعدين الآراء واستعرضنا المهام المختلفة التي تشكل جزءًا منها في العادة. كما عرضنا كيف يمكن استخدام الأدوات والأساليب التي ورد شرحها في الفصول السابقة (وبالأخص أدوات المعالجة اللغوية المسبقة وتمييز كيانات الأسماء وتمييز المصطلحات) جميعًا في مهمة تعدين الآراء، وكيفية بناء أداة من هذه الأدوات بدءًا من الصفر باستخدام هذه المكونات. هناك الكثير من التحديات التي لا تزال تعترض طريق عملية تطوير أدوات تعدين الآراء، ويبقى مستوى الأداء متدنيًا مقارنة بالعديد من مهام معالجة اللغات الطبيعية الأخرى، لكن هذا المجال يظل ميدانًا لعمليات البحث والتطوير التي تتم فيه على قدم وساق، على الرغم من أن الأدوات باتت تستخدم في سيناريوهات تجارية حقيقية. في الوقت الراهن، تسهم عملية دمج التقنيات الدلالية مثل البيانات المفتوحة المترابطة اللغوية (Linguistic Linked Open Data Cloud (LLOD)[1] السحابية مساهمة كبيرة في تحسين أداء هذه الأدوات وشموليتها، وفي الآونة الأخيرة برزت إمكانية أن تصبح أساليب التعلم العميق مجدية في مجال تعدين الآراء.

1- http://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/ (جرت زيارة الموقع في ٢٩ يناير ٢٠١٦).

الفصل الثامن

معالجة اللغات الطبيعية في شبكات التواصل الاجتماعي

تعد الاستفادة من الطابع الاجتماعي للتفاعلات التي تحدث بين البشر الركيزة الأساسية التي يقوم عليها انتشار وسائل التواصل الاجتماعي على نطاق واسع، وذلك من خلال تمكين الناس من التعبير عن آرائهم ولعب أدوار في مجتمع افتراضي والتعاون سويًّا عن بُعد. لو أخذنا التدوين القصير كمثال، يوجد في موقع تويتر أكثر من 300 مليون مستخدم نشط ينشرون ملايين التغريدات بشكل يومي(١).

في الوقت الراهن، بات التفاعل النشط مع هذه المسارات الإعلامية ذات القيمة العالية والأحجام الكبيرة ودورة الحياة القصيرة يمثل تحديًا يوميًّا يواجه المؤسسات والأفراد على حد سواء. ولذا فإن الحاجة لأتمتة هذه العملية بواسطة أساليب ذكية تعتمد على الدلالات للحصول على المعلومات باتت تتزايد بمرور الوقت. يمثل هذا الحقل ميدانًا جديدًا من ميادين البحث، ويجمع بين الأساليب المستخدمة في مجالات متعددة؛ مجال معالجة اللغات الطبيعية والعلوم الاجتماعية والتعلم الآلي والتشخيص واسترجاع المعلومات، بالإضافة إلى كونه يستخدم التقنيات الدلالية.

لم تعد أساليب البحث التقليدية قادرة على التعامل مع سلوكيات البحث عن المعلومات في شبكات التواصل الاجتماعي التي باتت أكثر تعقيدًا، فقد مرت تلك السلوكيات بعملية تحور سارت بها نحو صناعة المعنى (sense making) والتعلم والتحري والبحث الاجتماعي (social search) [215]. تملك التقنيات الدلالية إمكانات تتيح لها مساعدة البشر في التكيف بصورة أفضل مع المعلومات الفائضة الناتجة عن محتوى شبكات التواصل الاجتماعي. في نهاية المطاف، يمكن أن تسهم الأساليب الآلية المستندة إلى الدلالات والتي تتكيف مع أهداف الفرد في سعيه للحصول على المعلومات وتوفير ملخص موجز لمحتوى شبكات التواصل الاجتماعي ذي الصلة، في دعم عملية تفسير المعلومات وصناعة القرارات في ضوء موارد إعلامية واسعة النطاق وتتغير باستمرار.

وخلافًا للأخبار وغيرها من النصوص الموجودة على شبكة الإنترنت التي تجري صياغتها بعناية، تشكل موارد شبكات التواصل الاجتماعي عددًا من التحديات الماثلة

١- على سبيل المثال، تتراوح دقة أساليب تمييز كيانات الأسماء في العادة بين ٨٥٪ و٩٠٪ عندما تُطبق على المقالات الإخبارية، لكن دقتها تتراوح بين ٣٠٪ و٥٠٪ في التغريدات [٢١٩، ٢٢٠].

أمام تقنيات الدلالات، وذلك بسبب اتساع حجمها وطبيعتها المشوشة والعشوائية وكونها ذات طابع اجتماعي. يناقش هذا الفصل مهام معالجة اللغات الطبيعية وتحديات البحث التالية:

وجه الاختلاف بين تحليل شبكات التواصل الاجتماعي وغيرها من النصوص الطويلة الأقل تشويشا؛ الأنطولوجيات المطورة لنمذجة محتوى شبكات التواصل الاجتماعي ونتائج التحليل، وإضافة الشروح الدلالية إلى محتوى شبكات التواصل الاجتماعي مع التركيز على استخراج الكلمات/ المصطلحات الرئيسة، وتمييز كيانات الأسماء والربط بينها واستخراج الأحداث وتعدين المشاعر والآراء وإجراء تحليل مقارن لأنواع الوسائط المختلفة.

تمثل عملية البحث عن نتائج التحليل الدلالي لمحتوى شبكات التواصل الاجتماعي على نطاق واسع وتحويلها إلى صيغة صور مرئية مَهَمة في غاية الصعوبة، وهو ما سنناقشه في الفصل التاسع.

٨-١ مسارات شبكات التواصل الاجتماعي: الخصائص والتحديات والفرص

تتيح شبكات التواصل الاجتماعي للمستخدمين التواصل بعضهم مع بعض لغرض تبادل المحتوى (كروابط المواقع والصور ولقطات الفيديو) والتجارب والمعلومات المهنية، فضلاً عن التواصل مع الأصدقاء على الإنترنت. يقوم المستخدمون بإنشاء مشاركات أو تحديثات، وتقوم شبكات التواصل الاجتماعي بتعميمها على الدائرة الاجتماعية للمستخدم. الفرق الأساسي بين شبكات التواصل الاجتماعي وصفحات الويب التقليدية يكمن في أن مستخدمي شبكات التواصل الاجتماعي ليسوا مستهلكين غير فاعلين للمعلومات، بل يُعدُّ كثير منهم منتجين للمحتوى بغزارة.

يمكن تصنيف شبكات التواصل الاجتماعي حسب أطياف مختلفة أو بناءً على نوع التواصل بين المستخدمين أو وفق كيفية تبادل المعلومات أو طريقة تفاعل المستخدمين مع مسارات الوسائط:

تشجع وسائط رسم الاهتمامات (Interest-graph media) [216] مثل تويتر المستخدمين على إنشاء روابط مع المستخدمين الآخرين بناءً على اهتماماتهم المشتركة، بغض النظر عن كونهم يعرفون الشخص الآخر في الحياة العادية أم لا، ولا تتطلب الروابط دائمًا أن تتم من كلا الطرفين. تكون المعلومات المتبادلة على شكل مجموعة من الرسائل المعروضة وفق ترتيب زمني عكسي.

تشجع مواقع التواصل الاجتماعي (SNS) المستخدمين على التواصل مع الأشخاص الذين تجمعهم بهم علاقات حقيقية في الحياة العادية. يتيح موقع فيسبوك مثلاً طريقة لتبادل المعلومات بين الناس وإضافة التعليقات على مشاركات الآخرين. في العادة يجري تبادل مشاركات قصيرة ترسم صورة لمجريات حياة المستخدمين الحالية أو تتضمن رابطًا لأشياء موجودة على شبكة الإنترنت يعتقد المستخدم أن أصدقاءه قد يجدونها ممتعة. يجري جمع هذه التحديثات على شكل مجموعة مشاركات ذات ترتيب زمني يمكن لكل مستخدم قراءتها.

تهدف خدمات التواصل المهني (PNS) مثل لينكد إن (LinkedIn) إلى توفير خدمة تعارف في سياق مهني، حيث يعدُّ وجود رابط مع شخص معين بمنزلة شهادة تزكية منك لذلك الشخص إلى حد معين، وأنك توصي الآخرين بالعمل معه. في العادة يجري تبادل المعلومات المهنية عبر خدمات التواصل المهني التي تميل لاستقطاب المهنيين المتقدمين في العمر [217].

خدمات تبادل المحتوى والنقاش، كالمدونات ومواقع تبادل الفيديوهات (كيوتيوب وفيميو Vimeo) ومواقع تبادل العروض التقديمية (كموقع SlideShare) ومنتديات النقاش أو التقييم (مثل CNET). تتضمن المدونات في العادة مشاركات أطول، وبإمكان القراء التعليق عليها، كما تقوم بعض المدونات بإنشاء مقالات ذات تسلسل زمني ليطلع عليها القراء. تقوم العديد من المدونات أيضًا بالإعلان عن مستجدات مدوناتها بصورة آلية في حسابات مستخدميهم على فيسبوك وتويتر.

هذه الأنواع المختلفة من وسائل التواصل الاجتماعي، إلى جانب خصائصها المعقدة، تجعل عملية التفسير الدلالي شديدة الصعوبة. جرى تطوير الخوارزميات الحديثة التي تقوم بإضافة الشروح الدلالية وعمليات التصفح والبحث الآلي في المقام

الأول للمقالات الإخبارية وغيرها من أنواع المحتوى الإلكتروني التي تتميز بطولها وبجودة كتابتها [218]. على النقيض من ذلك، تعدُّ تحديثات المشاركات في وسائل التواصل الاجتماعي (كتغريدات تويتر ورسائل فيسبوك) متشابكة بقوة، وذات دورة حياة قصيرة، وهي مشوشة وقصيرة وتعج بالتعبيرات العامية، وهو ما يؤدي إلى نتائج رديئة للغاية[١].

تطرح هذه الخصائص -والتي تعد صعوبات في وسائل التواصل الاجتماعي- فرصًا أمام تطوير منهجيات جديدة في التقنيات القائمة على الدلالات تكون مناسبة بصورة كبرى لوسائل التواصل الاجتماعي:

الرسائل القصيرة (النصوص المصغّرة): تغريدات تويتر وغالبية رسائل فيسبوك قصيرة جدًّا (١٤٠ حرف للتغريدات). تعزز الكثير من الأساليب القائمة على الدلالات التي سنستعرضها أدناه هذه التغريدات والرسائل بمعلومات إضافية وسياق مأخوذ من الروابط المضمنة فيها والوسوم (الهاشتاغ)[٢]. على سبيل المثال، تعزز دراسة (أبيل وآخرون) [134] التغريدات من خلال ربطها بمقالات إخبارية صادرة في الحيز الزمني نفسه، في حين تستغل دراسة (مينديز وآخرون) قوائم علامات الوسوم الموجودة على الإنترنت لتعزيز التغريدات [221].

المحتوى المشوش: غالبًا ما يتضمن محتوى وسائل التواصل الاجتماعي أساليب غير مألوفة في التهجئة (مثال: 2moro [بدلاً من tomorrow])، واستخدام الأحرف الكبيرة بصورة غير منتظمة (مثال: تكبير أو تصغير جميع الأحرف) ورموز المشاعر (مثال: P-:)، والاختصارات التمييزية (مثال: ROFL وZOMG). تم تطوير أساليب لتحويل النص إلى الشكل القياسي [222]، بالإضافة إلى بعض الدراسات حول الاختلافات اللغوية القائمة على الموقع بين أنماط التقصير في النصوص المصغرة [223]. كما تُستخدم رموز المشاعر كمؤشرات مشاعر قوية في خوارزمية تعدين الآراء (راجع القسم ٨-٣-٤).

١- توصلت دراسة حديثة شملت ١,١ مليون تغريدة أن ٢٦٪ من التغريدات الإنجليزية تحتوي على عنوان URL فيما تحتوي ١٦,٦٪ من التغريدات علامة هاشتاغ، كما تتضمن ٥٤,٨٪ إشارة لاسم المستخدم [١٣٦].

2- http://xmlns.com/foaf/0.1/

الحيز الزمني: بالإضافة إلى التحليل اللغوي، محتوى وسائل التواصل الاجتماعي قد يناسب التحليل المعتمد على المسارات الزمنية، وهي مشكلة لم تحظ بقدر كافٍ من البحث. من الشروط الأساسية التي ينبغي توفرها في نماذج المعلومات المتعارضة والمتوافقة التي نحن بأمس الحاجة إليها التعامل مع مسألة كون وسائل التواصل الاجتماعي ذات حيز زمني مؤقت، بالإضافة إلى نمذجة التغيير في اهتمامات المستخدمين. علاوة على ذلك، يمكن دمج النمذجة الزمنية مع تعدين الآراء من أجل معاينة درجة التقلب في المواقف تجاه الموضوعات مع مرور الوقت.

السياق الاجتماعي: مهم لتفسير محتوى وسائل التواصل الاجتماعي بصورة صحيحة. كما ينبغي أن تستغل الأساليب القائمة على الدلالات سياق وسائل التواصل الاجتماعي (مثال: من الشخص الذي يتواصل معه المستخدم حاليًّا، وكم عدد مرات التواصل بينهم)، من أجل اشتقاق نماذج دلالية بصورة آلية لشبكات التواصل الاجتماعي وقياس سلطة المستخدم وتجميع المستخدمين المتشابهين ضمن مجموعات، فضلاً عن إيجاد نموذج يعكس مدى موثوقية العلاقة بين الطرفين ومتانتها.

المحتوى الناتج عن المستخدم: بالنظر لكون المستخدمين يقومون بإنتاج محتوى شبكات التواصل الاجتماعي وكذلك استهلاكها، هناك مصدر غني بالمعلومات الصريحة والمعلومات الضمنية المتعلقة بالمستخدم، بما في ذلك المعلومات الديموغرافية (الجنس، الموقع، العمر، ...الخ) والاهتمامات والآراء. يتمثل التحدي هنا في أن المحتوى الناتج عن المستخدم يكون محدودًا نسبيًّا في بعض الحالات؛ لذا لا يمكن تطبيق الأساليب الإحصائية المستندة إلى المكانز عليه بصورة ناجحة.

تعدد اللغات: يتميز محتوى شبكات التواصل الاجتماعي بكونه متعدد اللغات بصورة كبيرة. فعلى سبيل المثال، تقل نسبة التغريدات التي تُنشر باللغة الإنجليزية عن ٥٠٪، فيما تحتل اللغات اليابانية والإسبانية والبرتغالية والألمانية موقعًا بارزًا [136]. لكن مما يؤسف له كون التقنيات الدلالية قد ركزت حتى الآن في أغلبها على اللغة الإنجليزية، في حين تبقى مسألة تعديلها لتتلاءم مع لغات جديدة لم تُحسم بعد. يعدُّ التمييز الآلي للغات [136، 224] خطوة أولى مهمة، حيث تسمح للتطبيقات بالتمييز أولاً بين أنواع محتوى شبكات التواصل الاجتماعي وفقًا لمجموعات لغوية يمكن معالجتها بعد ذلك باستخدام خوارزميات مختلفة.

الجدول ٨-١ : الأنطولوجيات وما تقوم بنمذجته

الأنطولوجيا	الأشخاص	المشاركات الإلكترونية	شبكات التواصل الاجتماعي	المدونات المصغَّرة	اهتمامات المستخدمين	بطاقات التصنيف (Tags)	الموقع الجغرافي	سلوك المستخدم
FOAF	نعم		يعرف knows		جزئياً partial			
SIOC(T)	نعم	نعم		جزئياً	نعم			
MOAT						نعم		
Bottari	نعم	نعم	نعم	نعم		نعم	نعم	
DLPO	نعم	نعم	نعم	نعم	نعم	نعم		
SWUM	نعم				نعم		نعم	نعم
UBO	نعم		نعم		نعم			نعم

تناقش باقي أقسام هذا الفصل كيف يتم التعامل مع هذه التحديات في الأعمال البحثية التي أجريت حتى الآن، ونتطرق إلى بعض الجوانب التي ما زالت تعد قضايا مطروحة للنقاش.

٨-٢ استخدام الأنطولوجيات لتمثيل دلالات وسائل التواصل الاجتماعي

تُستخدم الأنطولوجيات بكثافة في عملية إضافة الشروح الدلالية وغيرها من أدوات معالجة اللغات الطبيعية. ونتيجة لذلك، سوف نركز في هذا القسم على الأنطولوجيات على وجه التحديد، فالأنطولوجيات يمكن أن تساعد أساليب معالجة اللغات الطبيعية فيما يتعلق بمختلف وسائل التواصل الاجتماعي والمحتوى المصاحب لها، بما في ذلك ملفات المستخدمين والمشاركات ووضع علامات التصنيف وإضافة الروابط. يعرض الجدول ٨-١ نظرة عامة على هذه الأنطولوجيات، إضافة إلى الجوانب المختلفة التي سيرد نقاشها بالتفصيل في القسم التالي:

شرح الأشخاص وشبكات التواصل الاجتماعي: مصطلحات صديق - لصديق[١] (FOAF Friend-of-a-Friend) هي مجموعة مصطلحات تُستخدم لوصف الأشخاص، حيث يضم الوصف أسماء الأشخاص وبيانات الاتصال وعلاقة معرفة (knows) عمومية. كما تدعم مصطلحات FOAF إمكانية النمذجة المحدودة للاهتمامات من خلال نمذجتها كصفحات على موضوعات الاهتمام. وكما تقر وثائق

1- http://sioc-project.org/

مصطلحات FOAF ذاتها، فإن مثل هذا النموذج الأنطولوجي الخاص بالاهتمامات محدود نوعًا ما.

نمذجة مواقع شبكات التواصل الاجتماعي: تقوم أنطولوجيا المجتمعات الإلكترونية المترابطة دلاليًّا[1] (SIOC) بنمذجة مواقع شبكات التواصل الاجتماعي (كالمدونات ومواقع الويكي والمنتديات الإلكترونية). تشمل المفاهيم الأساسية المنتديات والمواقع والمشاركات وحسابات المستخدمين ومجموعات المستخدمين وعلامات التصنيف. تدعم أنطولوجيا SIOC نمذجة اهتمامات المستخدمين بواسطة خاصية sioc :topic التي تكون قيمتها عبارة عن معرّف موارد موحد (URI) (كما أن المشاركات ومجموعات المستخدمين كذلك تحوي عناوين).

نمذجة المدونات المصغّرة: يوجد في أنطولوجيا SIOC امتدادات ظهرت في الآونة الأخيرة (SIOCT)، حيث تقوم هذه الامتدادات بنمذجة المدونات المصغرة باستخدام مفهوم MicroblogPost الجديد، وخاصية sioc : follows (التي تمثل العلاقات القائمة بين المتابعين والأشخاص الذين يتابعونهم على تويتر)، وخاصية sioc : addressed_to للمشاركات التي تذكر مستخدمين بعينهم. أنطولوجيا Bottari [225] هي أنطولوجيا جرى تطويرها خصيصًا لنمذجة العلاقات القائمة على موقع تويتر، ولا سيّما ربط التغريدات والمواقع ومشاعر المستخدمين (سواء أكانت إيجابية أم سلبية أم محايدة)، كامتدادات لأنطولوجيا SOIC. كما استُحدثت فئة جديدة تسمى TwitterUser، بالإضافة إلى خاصيتين منفصلتين هما follower وfollowing تشبهان الخصائص الموجودة في SIOCT. تنتمي فئة Tweet إلى النوع sioc:Post، وخلافًا لأنطولوجيا SIOCT، تميز أنطولوجيا Bottari أيضًا بين التغريدات المكررة والإجابات. كما يتم تمثيل المواقع بواسطة مصطلحات W3C الجغرافية[2]، وهو ما يتيح إمكانية إجراء التعليل المستند إلى المواقع.

الترابط بين وسائل التواصل الاجتماعي والشبكات الاجتماعية وممارسات المشاركات الإلكترونية: توفر أنطولوجيا DLPO نموذجًا شاملاً لمشاركات وسائل

1- http://www.w3.org/2003/01/geo/

2- http://www.w3.org/2004/02/skos/، طورت لنمذجة قواميس وقوائم مصطلحات ومصطلحات متحكم فيها.

التواصل الاجتماعي يتجاوز نطاق موقع تويتر [226]. كما أن لها جذورًا راسخة في الأنطولوجيات الأساسية كأنطولوجيا FOAF وأنطولوجيا SOIC ونظام ترتيب المعلومات البسيط (SKOS)(1). تقوم أنطولوجيا DLPO بنمذجة المعرفة الشخصية والاجتماعية المكتشفة من وسائل التواصل الاجتماعي، بالإضافة إلى ربط المشاركات عبر الشبكات الاجتماعية الشخصية. كما تضم هذه الأنطولوجيا ستة أنواع رئيسة من المعرفة، وهي المشاركات الإلكترونية وأنواع المشاركات المختلفة (كالتغريدات المكررة) والمشاركات المصغرة والحضور الإلكتروني (online presence) والحضور المادي وممارسات المشاركات الإلكترونية (كاستخدام الروابط والإضافة إلى قائمة التفضيلات). غير أنه على الرغم من أن الموضوعات والكيانات والأحداث وكذلك الأزمان قد نالت حظها من النقاش، إلا أن أدوار سلوك المستخدم والسمات الشخصية لم تُعالج بصورة شاملة في أنطولوجيا SWUM [227] التي يرد نقاشها أدناه.

نمذجة دلالات علامات التصنيف: تسمح أنطولوجيا MOAT (وهي اختصار لـMeaning-Of-A-Tag (معنى علامة التصنيف)) [228] للمستخدمين تحديد المعنى الدلالي لعلامات التصنيف من خلال ربط البيانات المفتوحة وإنشاء شروح دلالية لوسائل التواصل الاجتماعي في نهاية المطاف. تحدد هذه الأنطولوجيا تعريف اثنين من علامات التصنيف، وهما علامة التصنيف العمومية (أي تشمل المحتوى بأكمله) وعلامات التصنيف المحلية (علامات تصنيف خاصة بمصدر معين). يمكن دمج أنطولوجيا MOAT مع أنطولوجيا SIOCT من أجل تصنيف مشاركات المدونات المصغرة [229]. كما تقوم أنطولوجيا DLPO التي ورد شرحها أعلاه بنمذجة الموضوعات وعلامات التصنيف المرتبطة بالمشاركات الإلكترونية (بما في ذلك المدونات المصغرة).

أنطولوجيات نمذجة المستخدم مهمة لتمثيل معلومات المستخدمين وتفاعلاتهم على وسائل التواصل الاجتماعي وتجميعها ومشاركتها. على سبيل المثال، تهدف أنطولوجيا نمذجة المستخدم العمومية (GUMO) [230] إلى تغطية نطاق واسع من معلومات المستخدمين كالبيانات الديموغرافية وبيانات الاتصال وأنواع الشخصيات ...الخ.

1- http://twittersentiment.appspot.com/

غير أنها لا ترقى إلى مستوى تمثيل اهتمامات المستخدمين، وهو ما يجعلها غير ملائمة لوسائل التواصل الاجتماعي.

بناء على تحليل أجري على ١٧ تطبيقًا اجتماعيًّا من تطبيقات الشبكات الاجتماعية، قام (بلومباوم وآخرون) [227] باشتقاق عدد من أبعاد نموذج المستخدم المطلوبة لبناء أنطولوجيا نمذجة مستخدمي الشبكات الاجتماعية. تشمل تصنيفات الأبعاد التي اعتمدوها المعلومات الديموغرافية والاهتمامات والتفضيلات والاحتياجات والأهداف والحالة العقلية والجسدية والمعرفة والخلفية وسلوك المستخدم والسياق والسمات الشخصية (كالنمط الإدراكي ونوع الشخصية). وبناء على تلك الأمور، قاموا بإنشاء أنطولوجيا SWUM (نموذج مستخدم الويب الاجتماعي). لكن من عيوب أنطولوجيا SWUM عدم اعتمادها على الأنطولوجيات الأخرى. على سبيل المثال، يتم ترميز خصائص موقع المستخدم كالبلد والمدينة على شكل تسلسلات (strings)، وهو ما يحد بشكل كبير من جدواها في مجال التعليل (مثال: من الصعب إيجاد جميع المستخدمين المتواجدين في جنوب غرب إنجلترا، بالاعتماد على مدنهم). تتمثل المنهجية البديلة التي يمكن استخدامها في تحديد تعريف تلك الخصائص بواسطة معرف الموارد الموحد (URI) الذي يرتكز على موارد البيانات المترابطة (Linked Data) التي يشيع استخدامها، مثل DBpedia وFreebase.

أخيرًا، تقوم أنطولوجيا سلوك المستخدم [231] بنمذجة تفاعلات المستخدمين في المجتمعات الإلكترونية. كما جرى استخدامها لنمذجة سلوك المستخدم في المنتديات الإلكترونية [231] وكذلك النقاشات على تويتر [232]. يوجد فيها أيضًا فئات (classes) تقوم بنمذجة تأثير المشاركات (الإجابات والتعليقات ...الخ) وسلوك المستخدم وأدوار المستخدم (على سبيل المثال: مُبادر ذو شعبية، داعم، مُهمَل) والسياق الزمني (الإطار الزمني) وغيرها من معلومات التفاعل. تحظى مسألة معالجة البعد الزمني لوسائل التواصل الاجتماعي بأهمية خاصة، ولا سيّما عند نمذجة التغييرات التي تحدث بمرور الوقت (كالتغييرات التي تؤثِّر في اهتمامات المستخدمين وآرائهم).

وكتلخيص لما سبق، هناك عدد من الأنطولوجيات المتخصصة التي تهدف إلى تمثيل المعلومات الدلالية المشتقة بصورة آلية من وسائل التواصل الاجتماعي وتعليلها. غير

أنه بالنظر إلى كونها تعالج ظواهر مختلفة، فإن تطبيقات معالجة اللغات الطبيعية تعتمد أكثر من أنطولوجيا واحدة أو توسّع نطاق عملها لتلبية متطلباتها. في بعض الحالات، يجري استخدام أساليب معالجة اللغات الطبيعية لتعبئة هذه الأنطولوجيات بالحالات (instances) بصورة تلقائية، وذلك استنادًا إلى محتوى وسائل التواصل الاجتماعي (مثل تعبئة نماذج المستخدمين والمجتمعات الخاصة بمجموعة محددة من المستخدمين/ المجتمعات).

٨-٣ إضافة الشروح الدلالية إلى وسائل التواصل الاجتماعي

قام الباحثون بالتحقيق في مجموعة كبيرة من مهام إضافة الشروح الدلالية إلى محتوى وسائل التواصل الاجتماعي. يناقش هذا القسم جانبًا من هذه الأمور بمزيد من التفصيل، بداية من مهمة استخراج العبارات المفتاحية.

٨-٣-١ استخراج العبارات المفتاحية

تتميز العبارات المفتاحية المختارة بصورة آلية بكونها مفيدة في تمثيل موضوع وثيقة معينة أو مجموعة من الوثائق، على الرغم من أنها ليست فعالة جدًّا في عرض الحجج أو الإفادات الكاملة الموجودة في تلك الوثائق. لذلك يمكن اعتبار استخراج العبارات المفتاحية نوعًا من استخراج المعرفة السطحي الذي يقدم لمحة عامة موضعية. وفي سياق إضافة الشروح الدلالية واسترجاعها، يمكن استخدام الكلمات المفتاحية أيضًا كأداة لتقليل تعدد الأبعاد (dimensionality) والسماح للنظام بالتعامل مع مجموعة أقل من المصطلحات المهمة بدلاً من الوثيقة بأكملها.

تعدُّ مهمة استخراج العبارات المفتاحية وثيقة الصلة بمهمة استخراج المصطلحات، إلا أنها تختلف عنها في المقام الأول في كونها ذات طابع تمثيلي. تهدف مهمة استخراج العبارات المفتاحية إلى تمثيل الموضوع عن طريق استخراج الكلمات والعبارات الأكثر أهمية، ولذا فهي تعطي نظرة عامة نوعًا ما عن الوثيقة، ولذا فإن لديها هدفًا نهائيًّا واضحًا. من جهة أخرى لا تسعى مهمة استخراج المصطلحات إلى تمثيل الوثيقة بصورة مباشرة، لكنها تحاول فقط إيجاد المصطلحات ذات النطاق المحدد (domain-specific) التي جرى استخدامها (بغض النظر عن مدى أهمية تلك المصطلحات بالوثيقة نفسها). أيضًا

في حين تكون المصطلحات المستخرجة مرتبطة بأنطولوجيا أو بمصطلحات أخرى، هذا الأمر لا ينطبق على عملية استخراج العبارات المفتاحية.

تستغل بعض منهجيات استخراج الكلمات المفتاحية التوارد المشترك بين المصطلحات (term co-reference)، إذ تقوم بإنشاء رسم بياني مكون من مصطلحات وله حواف (edges) مشتقة من المسافة الفاصلة بين أزواج المصطلحات الواردة في النص، وإعطاء أوزان لزوايا (vertices) الرسم البياني [233]. أنشئ هذا النوع من استخراج الكلمات المفتاحية للحصول على أداء جيد عند معالجة بيانات تويتر مقارنة بالأساليب المستندة إلى نماذج النصوص [234].

ولعل من أسباب الأداء الجيد الذي تقدمه المنهجيات المستندة إلى الرسوم البيانية المستخدمة في استخراج الكلمات المفتاحية من تويتر كونَ هذا النطاق يحتوي على قدر كبير من التكرار [235]. على سبيل المثال، في سياق الموضوعات الأكثر تداولاً على تويتر (التي يُشار إليها بواسطة علامات الهاشتاغ)، قامت دراسة [236] باستخراج عبارات مفتاحية عن طريق الاستفادة من التكرار النصي واختيار التسلسلات الشائعة للكلمات. وفي حين يعدُّ التكرار في تويتر وغيره من شبكات التواصل الاجتماعي مفيدًا نوعًا ما عندما يتعلق الأمر بإنشاء ملخصات الكلمات المفتاحية، هناك سمة أخرى أقل فائدة، وهي التنوع الكبير في الموضوعات التي تجري مناقشتها. في الحالات التي تناقش فيها الوثائق أكثر من موضوع واحد، قد تكون هناك صعوبة في استخراج مجموعة متناسقة ودقيقة من الكلمات منها.

عند التعامل مع تحديثات توتير الشخصية على أنها وثيقة واحدة، فإنها تطرح هذه الإشكالية. بصورة عامة، يستطيع المستخدمون نشر مشاركات تتناول عدة موضوعات. وفي حين تستخدم دراسة [234] أداة TextRank لمعالجة جميع تحديثات المستخدم، إلا أن الباحثين في تلك الدراسة لم يحاولوا نمذجة التباين في الموضوعات أو التعامل معه، وذلك على عكس الباحثين في دراسة [237] الذين قاموا بدمج مهمة نمذجة الموضوعات في منهجيتهم. لم تكن دراستهم الدراسة الوحيدة التي قامت بتطبيق نمذجة الموضوعات على بيانات تويتر، وذلك لأن دراسة [238] قامت بذلك أيضًا. غير أنه في الدراسة الأخيرة لم يجرِ تلخيص الموضوعات على الرغم من استكشافها.

في سياق خدمات التصنيفات والتفضيلات الاجتماعية مثل Flickr وDelicious وBibsonomy، درس الباحثون التصنيف التلقائي للوثائق الجديدة بواسطة بطاقات التصنيف (tags) الخاصة بالفهرسة الجماعية (folksonomy). يعدُّ نظام AutoTag من أوائل المنهجيات [239]، حيث يقوم هذا النظام بإضافة بطاقات تصنيف إلى مشاركات المدونات. في البداية، يعثر النظام على مدونات متشابهة ومفهرسة مسبقًا باستخدام أساليب استرجاع المعلومات المعيارية، وذلك باستخدام المدونة الجديدة كاستفسار. بعد ذلك يقوم بإنشاء قائمة مرتبة مكونة من بطاقات تصنيف (tags) مأخوذة من المشاركات الأكثر صلة، ومعززة بمعلومات عن بطاقات التصنيف التي استخدمها صاحب المدونة المعنيّ.

تستخدم المنهجيات الحديثة عملية استخراج العبارات المفتاحية من محتوى المدونات من أجل اقتراح بطاقات تصنيف جديدة. على سبيل المثال، تقوم دراسة [240] بتوليد عبارات مفتاحية محتملة من سلاسل ن-جرام (n-grams)، وذلك اعتمادًا على بطاقات تصنيف أقسام الكلام (POS) الخاصة بها، وبعدها تقوم بفرزها باستخدام مُصنِّف انحدار لوجستي (logistic regression classifier). يمكن دمج الأسلوب القائم على العبارات المفتاحية مع المعلومات المستمدة من الفهرسة الجماعية (folksonomy) [241]، وذلك من أجل توليد توقيعات بطاقات التصنيف (tag signatures) (أي ربط كل بطاقة تصنيف في الفهرسة الجماعية بمصطلحات موزونة ومترابطة دلاليًا). بعد ذلك تجري المقارنة بينها وترتيبها في ضوء المدونة الجديدة، وذلك من أجل اقتراح بطاقات التصنيف الأكثر صلة.

٨-٣-٢ تمييز كيانات الأسماء المستند إلى الأنطولوجيات في وسائل التواصل الاجتماعي

ثبت أن أساليب تمييز كيانات الأسماء، التي يجري تدريبها عادة على النصوص الطويلة الأكثر انتظامًا (كالمقالات الإخبارية) تعطي أداءً سيئًا عند تطبيقها على محتوى وسائل التواصل الاجتماعي التي تتسم بكونها أقصر وأكثر تشويشًا من أنواع المحتوى الأخرى [220]. غير أنه في حين تقدم كل مشاركة على حدة سياقًا لغويًّا غير مكتمل، إلا أنه يمكن الحصول على معلومات إضافية من ملفات المستخدمين وشبكات التواصل الاجتماعي والمشاركات المترابطة (كالردود على رسائل التغريدات). يناقش هذا القسم ما نسميه

منهجيات إضافة التعليقات والشروح الدلالية الموجهة لوسائل التواصل الاجتماعي، التي تدمج بين السمات اللغوية والسمات الخاصة بوسائل التواصل الاجتماعي.

يتناول (ريتر وآخرون) في دراسة [220] مشكلة تصنيف كيانات الأسماء (لكن ليس إزالة الغموض عنها) باستخدام قاعدة المعرفة Freebase كمصدر لعدد كبير من الكيانات المعروفة. من دون أخذ السياق بعين الاعتبار، لا يحقق النظام المبسط للبحث عن الكيانات وتحديد النوع سوى نسبة ٣٨٪ في درجة f (f-score) (تكون ٣٥٪ من الكيانات غامضة ولديها أكثر من نوع واحد، في حين لا تظهر ٣٠٪ من الكيانات الموجودة في التغريدات في قاعدة المعرفة Freebase). عند تطبيق تصنيف كيانات الأسماء يتحسن الأداء ليصل إلى ٦٦٪، وذلك عبر استخدام نماذج موضوعات مصنفة تأخذ السياق بعين الاعتبار وكذلك التوزيع على أنواع Freebase لكل تسلسل من تسلسلات الكيانات (مثال: يمكن أن تكون أمازون شركة أو موقعًا).

تتناول دراسة (آيرسون وآخرون) [242] مشكلة إزالة الغموض (تحديد أسماء المواقع الجغرافية) عن موقع بطاقات التصنيف في Flickr. تقوم هذه المنهجية على أساس قاعدة المعلومات الدلالية GeoPlanet التابعة لياهو، حيث تقوم بإعطاء معرّف موارد موحد (URI) لموقع كل حالة (instance)، بالإضافة إلى تصنيف مكوّن من مواقع مترابطة (مثال: المواقع المتجاورة). تستخدم منهجية إزالة الغموض عن بطاقات التصنيف جميع بطاقات التصنيف الأخرى المعطاة للصورة، وكذلك سياق المستخدم (جميع بطاقات التصنيف المعطاة من قبل هذا المستخدم لجميع الصور الخاصة به)، وسياق المستخدم الممتد الذي يأخذ بعين الاعتبار بطاقات التصنيف الخاصة بجهات الاتصال الموجودة لدى المستخدم. وقد جرى إثبات أن استخدام هذا السياق الأوسع المعتمد على الدائرة الاجتماعية يحسن بشكل كبير دقة عملية إزالة الغموض بصورة عامة.

هناك مصدر آخر للدلالات الإضافية الضمنية، وهي علامات الهاشتاغ المستخدمة في رسائل تويتر، التي تحولت إلى وسيلة تتيح للمستخدمين متابعة النقاشات الدائرة حول موضوع معين. قام لانيادو وميكا [243] بالتحقيق في دلالات علامات الهاشتاغ في ٣٦٩ مليون رسالة، مستخدمين أربعة مقاييس هي تكرار الاستخدام، ودرجة التحديد (استخدام علامات الهاشتاغ بدل كلمة ما في مقابل استخدام الكلمة نفسها)،

وتناسق الاستخدام، والثبات بمرور الوقت. بعد ذلك تُستخدم تلك المقاييس لتحديد علامات الهاشتاغ التي يمكن استخدامها كمُعرّفات ومن ثمّ تُربط بمُعرّفات الموارد الموحدة (URI) الخاصة بقاعدة معلومات Freebase (معظمها عبارة عن كيانات أسماء). استُخدمت علامات الهاشتاغ أيضًا كمصدر إضافي للمعلومات الدلالية المتعلقة بالتغريدات، وذلك بإضافة تعريفات نصية لعلامات هاشتاغ مأخوذة من قواميس إلكترونية جماهيرية [221]. بدورهم قام (ميينديز وآخرون) [221] بإضافة الشروح الدلالية عن طريق إجراء بحث بسيط عن الكيانات مقارنة بالكيانات والفئات الموجودة في DBpedia من دون إزالة الغموض بصورة كبرى. جرى ترميز الخصائص ذات الصلة بالمستخدم وكذلك الارتباطات الاجتماعية في FOAF، بينما جرى ترميز الشروح الدلالية في أنطولوجيا MOAT (راجع القسم ٨-٢).

الجدول ٢, ٨: إضافة الشروح الدلالية بواسطة الأنطولوجيات: أدوات بحث مختارة

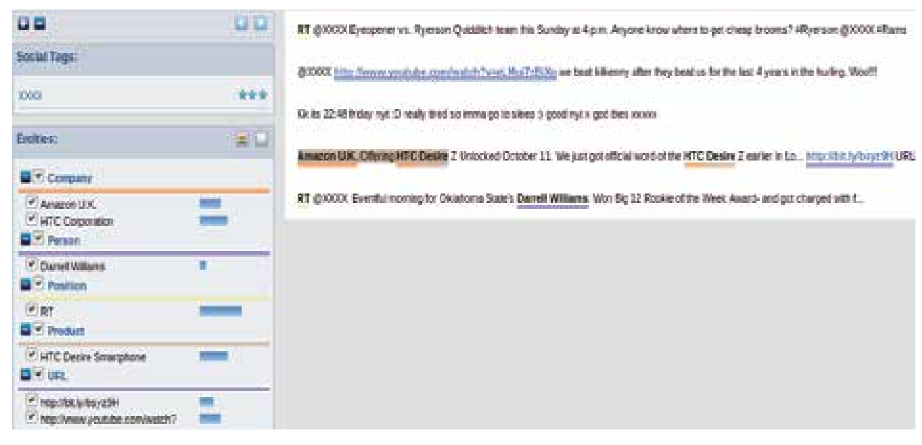
	الأنطولوجيا/ مورد البيانات المفتوحة المترابطة المستخدم	الشروحات الناتجة	إزالة الغموض	النطاق المستهدف	المكنز المستخدم	التقييم بواسطة
DBpedia Spotlight [115]	DBpedia, Freebase	أكثر من ٣٠ نوع	نعم	نطاق مفتوح	ويكيبيديا	الأخبار
LINDEN [117]	YAGO	أنواع YAGO	نعم	نطاق مفتوح	ويكيبيديا	TAC-KBP 2009
Ritter [220]	Freebase	١٠ أنواع	لا	نطاق مفتوح	تغريدات	تغريدات
Ireson [242]	GeoPlanet	المواقع	نعم	الصور	فليكر	فليكر
Laniado & Mika [243]	Freebase	Freebase	نعم	نطاق مفتوح	التغريدات	تغريدات

	الأنطولوجيا/ مورد البيانات المفتوحة المترابطة المستخدم	الشروحات الناتجة	إزالة الغموض	النطاق المستهدف	المكنز المستخدم	التقييم بواسطة
Meij [121]	Wikipedia	Wikipedia	نعم	نطاق مفتوح	ويكيبيديا	تغريدات
Gruhl [244]	MusicBrainz	الأغاني والألبومات	نعم	نطاق الموسيقى	مايسبيس	مشاركات مايسبيس
Rowe [134]	DBpedia	ذات صلة بالتوارد المشترك	نعم	المؤتمرات	تغريدات	٢٠٠ تغريدة
Choudhury [245]	Wikipedia	لاعبو الكريكت، الألعاب	نعم	الأحداث الرياضية	ويكيبيديا	تغريدات الكريكت

تستفيد منهجيات ربط الكيانات المستندة إلى ويكيبيديا (راجع القسم ٣,٥) بصورة كبيرة من السياق اللغوي الأكبر المتوفر في المقالات الإخبارية وصفحات الويب. قدَّم تقييم DBpedia Spotlight [115] وطريقة Milne وWitten [114] باستخدام قاعدة بيانات مكونة من تغريدات أداءً أسوأ بكثير [121]. يقترح (ميج وآخرون) [121] استخدام منهجية خاصة بتويتر لربط هذا النوع من الرسائل القصيرة والمشوشة بمقالات ويكيبيديا. تستخدم الخطوة الأولى سلاسل ن-جرام (n-grams) لتوليد قائمة من مفاهيم ويكيبيديا المحتملة، وبعد ذلك يُستخدم أسلوب التعلم الخاضع للإشراف لتصنيف كل مفهوم على أنه إما مفهومٌ ذو صلة أو مفهوم غير ذي صلة (في ضوء التغريدة والمستخدِم الذي قام بكتابتها). تستخدم هذه الوسيلة خصائص مستمدة من سلاسل ن-جرام (n-grams) (كعدد مقالات ويكيبيديا التي تضم سلسلة ن-جرام هذه)، وخصائص مقالات ويكيبيديا (كعدد المقالات التي تحتوي على رابط للصفحة المعنية)، وخصائص التغريدات (كاستخدام تعريفات علامات الهاشتاغ وصفحات الويب المترابطة).

يركز (جروهل وآخرون) [244] بصفة خاصة على عنصر إزالة الغموض في عملية إضافة الشروح الدلالية ويقومون بدراسة مشكلة التعامل مع الحالات شديدة الغموض، مثلما هو الحال مع عنوانات الأغاني والألبومات الموسيقية. تقوم المنهجية التي يعتمدونها أولاً بتقييد الجزء الموجود في أنطولوجيا MusicBrainz المستخدم لإنتاج الاحتمالات (في هذه الحالة يكون ذلك عن طريق إزالة جميع المعلومات المتعلقة بالفنانين الموسيقيين الذين لم يرد ذكرهم في النص المعني). ثانيًا، يقومون بتطبيق مهام معالجة اللغات السطحية، مثل تصنيف أقسام الكلام وتجزئة العبارات الاسمية، وبعد ذلك يستخدمون هذه المعلومات كمُدخلات لُمصنّف آلة دعم المتجه (support vector machine classifier) والذي بدوره يقوم بإزالة الغموض بناءً على أساس هذه المعلومات. اختُبرت هذه المنهجية على مكنز يضم مشاركات موقع MySpace لثلاثة فنانين. وعلى الرغم من أن الأنطولوجيا كبيرة جدًّا (الأمر الذي يولد الكثير من الغموض)، إلا أن النصوص شديدة التركيز، وهو ما يسمح للنظام بتحقيق أداء جيد. وكما ذكر القائمون على الدراسة أنفسهم، من المرجح أن تطرح عملية معالجة النصوص الأقل تركيزًا كرسائل تويتر أو المقالات الإخبارية تحديًا أكبر بكثير.

وفيما يتعلق بربط الكيانات، كشفت التقييمات التي تناولت تغريدات تويتر في الآونة الأخيرة عن وجود مشكلات في استخدام المنهجيات العصرية في هذا النوع [67، 134]، ويعود سبب ذلك إلى حد بعيد إلى قِصر التغريدات (١٤٠ حرفًا) وأيضًا إلى التعامل مع كل مشاركة على حدة من دون أخذ السياق الأشمل المتاح بعين الاعتبار. على وجه الخصوص، تجري معالجة نصوص التغريدات فقط في العادة، على الرغم من أن عنصر JSON (JSON object) يضم أيضًا بيانات تتعلق بملف المستخدم (الاسم الكامل، الموقع الاختياري، نص الملف الشخصي، وصفحة الويب). كما تشمل قرابة ٢٦٪ من جميع التغريدات عنوانات URL [136] وتضم ١٦,٦٪ منها علامات هاشتاغ، في حين تحتوي ٥٤,٨٪ منها إشارة إلى اسم مستخدم واحد على الأقل.



الشكل ٨-١: نتائج كاليه للتغريدات.

لا تستفيد أنظمة تمييز كيانات الأسماء التي تستهدف المدونات المصغّرة في العموم من إشارات وسائل التواصل الاجتماعي، فهي تتعامل مع علامات الهاشتاغ مثلاً على أنها من الأسماء المشتركة (common nouns)، على سبيل المثال نظام [219، 246]، أو لا تعدها كذلك، مثلما هو الحال في نظام TwiNER [247]. تستخدم دراسة (شين وآخرون) [139] تغريدات إضافية مأخوذة من تحديثات المستخدم للعثور على موضوعات خاصة بالمستخدم واستخدام تلك الموضوعات لتحسين عملية إزالة الغموض. تطرح دراسة (هوانج وآخرون) [140] صيغة موسّعة لعملية إزالة الغموض المستندة إلى الرسوم البيانية تستحدث «مسارات وصفية» (Meta Paths) تمثل السياق المستمد من التغريدات الأخرى عبر علامات الهاشتاغ المشتركة أو مؤلفي التغريدات المشتركين أو الإشارات (mentions) المشتركة. تقوم دراسة (جاتاني وآخرون) [141] بتوسيع عنوانات URL المختصرة والسياق المستمد من التغريدات التي تعود إلى المؤلف نفسه والتغريدات التي تحتوي على علامات الهاشتاغ ذاتها، لكنها لا تُقيّم مساهمة هذا السياق الإضافي في الأداء النهائي، ولا تستغل تعريفات علامات الهاشتاغ كما لا تستخدم نصوص ملفات المستخدمين الشخصية.

في سياق نظام YODIE (راجع القسم ٢,٣,٥)، قام [129] بإجراء تحقيق ممنهج لدراسة تأثير السياق الاجتماعي الأشمل على أداء عمليات إزالة الغموض في التغريدات المستندة إلى البيانات المفتوحة المترابطة (LOD). وعلى وجه الخصوص ما يتعلق

بعلامات الهاشتاغ. جرى تعزيز محتوى التغريدات بتعريفات علامات الهاشتاغ المستمدة تلقائيًّا من شبكة الإنترنت. وبالمثل، جرى تعزيز التغريدات التي تحتوي على إشارات mentions@ بمعلومات نصية مستمدة من ذلك الملف الشخصي الموجود على تويتر المشار إليه بإشارة mention@. وفيما يتعلق بعنوانات URL، أرفقت التغريدة بنص الويب التي تحتوي عليها الروابط. جرى قياس أداء عملية إزالة الغموض في حالتين هما عند إجراء عملية توسيع النطاق بصورة فردية (أي استخدام علامات الهاشتاغ فقط، أو استخدام عنوانات URL فقط، ...الخ)، وكذلك عند استخدام جميع أنواع المعلومات السياقية مجتمعة. أظهرت الاختبارات أن توسيع التغريدات أدى إلى تحسن كبير في أداء عملية ربط الكيانات في محتوى المدونات المصغّرة. على وجه الخصوص، تحسنت الدقة الإجمالية بنسبة ٧,٣ في المائة، علمًا أن الزيادة في الأداء كانت أقل بالنسبة لدرجة F1، حيث سجلت ٦,٢ في المائة.

معظم التحسن في الأداء نتج عن القدرة على إزالة غموض إشارات mentions@، حيث أخفقت عملية استخدام نصوص التغريدات فقط في التعرف على مرجع DBpedia (referent) الذي تشير إليه تلك الإشارات. يتمثل المساهم الرئيس إذًا في تحسّن الأداء في هذه الحالة في الاستدعاء. ينبغي أيضًا ملاحظة أنه حتى من دون توسيع نطاق الإشارات، فقد أدى توسيع عنوانات URL وعلامات الهاشتاغ إلى حدوث تحسينات كبيرة.

معالجة محتوى وسائل التواصل الاجتماعي بواسطة منصة GATE

نظرًا للطبيعة الصعبة لوسائل التواصل الاجتماعي (راجع القسم ٨)، فقد جرى تكييف أدوات المعالجة المسبقة وتمييز الكيانات الموجودة في منصة GATE (راجع الفصلين الثاني والثالث) لتلائم هذا النوع من المحتوى.

لهذا السبب، توفر منصة GATE مكونًا إضافيًّا يسمى TwitIE [248]- وهو نسخة مخصصة من أداة ANNIE صممت خصيصًا لمحتوى وسائل التواصل الاجتماعي، وجرى اختبارها على نطاق واسع في رسائل المدونات المصغرة. يتسم محتوى المدونات المصغرة في كونه متاحًا بسهولة على شكل تحديثات عامة ضخمة، كما يعدُّ هذا المحتوى

الأصعب من حيث المعالجة بواسطة أدوات IE العمومية، وذلك بسبب كونه ذا طابع موجز ومشوش، وأيضًا لانتشار المصطلحات العامية فيه وأشكال التعبيرات المتعارف عليها في تويتر.

يظهر الشكل ٨-٢ مراحل منظومة TwitIE ومكوناتها. تتوفر منظومة TwitIE كمكون إضافي في منصة GATE، ويلزم تحميلها لكي تظهر موارد المعالجة هذه داخل مطور GATE Developer. تظهر المكونات المستمدة من أداة ANNIE التي لم يطرأ عليها أي تعديلات باللون الأزرق، في حين تعدُّ المكونات الظاهرة باللون الأحمر مكونات جديدة وخاصة بوسائل التواصل الاجتماعي.

تتمثل الخطوة الأولى في تحديد اللغة، وهي مهمة تعتمد على نسخة من TextCat جرى تعديلها لتتناسب مع وسائل التواصل الاجتماعي [136]. وبسبب قِصر التغريدات، يفترض النظام أن كل تغريدة مكتوبة بلغة واحدة. تُحدَّد اللغات المستخدمة للتصنيف بواسطة ملف تكوين (configuration file) يتم توفيره كمعامل تهيئة (initialization parameter). عند إعطاء مجموعة من التغريدات المكتوبة بلغة جديدة، يمكن تدريب نظام TextCat TwitIE لدعم تلك اللغة الجديدة أيضًا. يجري ذلك باستخدام برنامج توليد البصمات (Fingerprint Generation PR)، المدرج في مكون Language Identification الإضافي في منصة GATE [249]. يقوم البرنامج بإنشاء بصمة جديدة من مكنز مكون من الوثائق.

يعدُّ مجزئ الجمل TwitIE نسخة معدلة من مجزئ الجمل الإنجليزي الخاص بأداة ANNIE، وهو مبني على نظام Rite لتجزئة الجمل [220]. يتعامل هذا المجزئ على وجه التحديد مع الاختصارات (مثل RT وROFL) وعنوانات URL كوحدة لغوية واحدة لكل اختصار. تكون علامات الهاشتاغ والإشارات (mentions) وحدتين لغويتين (أي وحدة لعلامة # ووحدة أخرى لكلمة nike في المثال الوارد أعلاه) بالإضافة إلى هاشتاغ (HashTag) أضيفت إليه التعليقات والحواشي بحيث يغطي كلا الجانبين الاثنين. يتم الحفاظ على الأحرف الكبيرة، لكن تضاف خاصية تتعلق بالتهجئة: عندما تكون جميع الأحرف كبيرة، وعند استخدام الأحرف الصغيرة، وعند استخدام الأحرف الكبيرة والصغيرة المختلطة. تتم معالجة الأحرف الصغيرة والرموز التعبيرية

(emoticons) في وحدات منفصلة نظرًا لعدم وجود حاجة إليها في العادة. بناءً على ذلك، تكون عملية التجزئة أسرع وأكثر شمولية، وكذلك أكثر ملاءمة لاحتياجات عملية تمييز كيانات الأسماء.

يتألف المعجم الجغرافي (gazetteer) من قوائم كالمدن والمؤسسات وأيام الأسبوع وما إلى ذلك. لا تقتصر القوائم على الكيانات فحسب، بل تشمل أيضًا أسماء المؤشرات المفيدة كتسميات الشركات المعتادة (مثال: «محدودة»)، والعناوين وما إلى ذلك. تحوّل قوائم المعاجم الجغرافية برمجيًّا إلى آلات الحالة المنتهية (finite state machines)، التي يمكن أن تتطابق مع الوحدات اللغوية النصية. في الوقت الحالي، تعيد أداة TwitIE استخدام قوائم ANNIE الجغرافية من دون إجراء أي تعديل.

أداة تقسيم الجمل هي عبارة عن سلسلة تعاقبية مكونة من محولات طاقة منتهية الحالات (finite-state transducers) تقوم بتجزئة النص إلى جُمل. هذه الوحدة ضرورية لمُصنّف أقسام الكلام. مرة أخرى، يُعاد استخدام مقسّم الجمل الخاص بمنصة ANNIE من دون إجراء تعديل، على الرغم من أنه عند معالجة التغريدات، يمكن استخدام نص التغريدة كجُملة واحدة فقط من دون إجراء المزيد من التحليل.

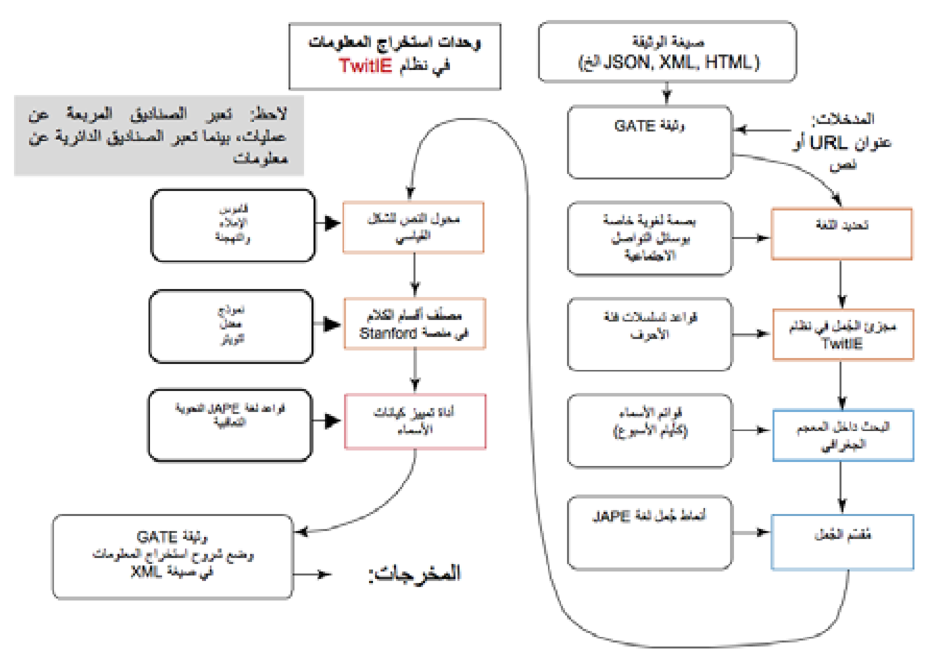
يعدُّ معيد النص إلى شكله القياسي في أداة TwitIE في الوقت الحالي مزيجًا بين قاموس عام لتصحيح الأخطاء الإملائية وقاموس آخر لتصحيح الأخطاء الإملائية خاص بوسائل التواصل الاجتماعية. يحتوي القاموس الأخير على مُدخلات (entries) من قبيل «2moro» و«brb»، مثلما هو الحال في دراسة (هان وآخرون) [250].

يضم مُصنّف أقسام الكلام نموذجًا معدلاً لمصنّف أقسام الكلام الخاص بمنصة Stanford، وهو مدرب على التغريدات المصنفة في مكنز Penn Treebank. أضيفت بطاقات تصنيف إضافية خاصة بالتغريدات المكررة وعنوانات URL وعلامات الهاشتاغ وإشارات المستخدمين (mentions)، كما أعيد تدريب مُصنّف Stanford لتصنيف أقسام الكلام [251] باستخدام بعض التغريدات التي أضيفت إليها الشروح يدويًّا [220]،

وكذلك مكنز NPS IRC [252] والنصوص الإخبارية (القسم الخاص بجريدة

وول ستريت جورنال في penn treebank [253]). يحقق النموذج الناتج دقة تصل إلى ١٤, ٨٣٪ في تصنيف أقسام الكلام، وهي نسبة تبقى دون نسبة الـ٩٧٪ التي تُحقق عند معالجة المحتوى الإخباري. لكي نضمن أفضل مستوى ممكن من الأداء، لا بد من تشغيل مُصنّف أقسام الكلام الخاص بأداة TwitIE بعد تشغيل محول النص إلى الشكل القياسي ومجزئ الجمل الموجودين في أداة TwitIE. ونظرًا لأنها تُدرّب في الوقت الحالي على المحتوى الإنجليزي فقط، من الضروري تشغيلها باستخدام التغريدات التي سبق تمييزها على أنها مكتوبة باللغة الإنجليزية بواسطة معرّف اللغات في أداة TwitIE.

أخيرًا، تكون وحدة تميز كيانات الأسماء في أداة TwitIE صيغة معدلة يدويًّا مقتبسة من أداة تميز الكيانات وفقًا للقواعد الخاصة بأداة ANNIE. وبفضل تعديل أداة ANNIE لملاءمة وسائل التواصل الاجتماعي، تحقق أداة TwitIE دقة مطلقة بنسبة تزيد على ٣٠٪ وزيادة في أداء F1 بنسبة ٢٠٪ مقارنة بأداة ANNIE.



الشكل ٨-٢: منظومة أداة TwitIE لاستخراج المعلومات.

٨-٣-٣ اكتشاف الأحداث

يمكن استخدام الكثير من الأشياء كالموضوعات الرائجة لمراقبة الآراء وردود الأفعال الدولية، كما يمكن استخدام تحديثات وسائل التواصل الاجتماعي كقناة خلفية تدور فيها النقاشات حول الأحداث التي تجري في العالم الحقيقي [254]، وكذلك لاكتشاف تلك الأحداث والإبلاغ عنها فور حدوثها تقريبًا. في حين قد يبدو للوهلة الأولى أن الموضوعات الرائجة وحدها تكفي لإنجاز هذه المهمة، إلا أن هناك عددًا من الأسباب التي تجعلها غير كافية:

العمومية: قد تتناول الموضوعات الرائجة ما يجري من أحداث، إلا أنها قد تشير أيضًا إلى المشاهير أو المنتجات أو الميمات الإلكترونية (online memes).

النطاق: الموضوعات التي تتفاعل معها شريحة عريضة من مستخدمي تويتر يمكن أن تظهر ضمن الموضوعات الرائجة دون غيرها.

الرقابة: يعتقد الكثيرون أن الموضوعات الرائجة المعروضة من قبل خدمة تويتر الرسمية تخضع للرقابة السياسية واللغوية.

الخوارزميات: الأسلوب المستخدم لاختيار الموضوعات الرائجة لا يُنشر في أي مكان وليس مفهومًا بصورة عامة.

إذًا يطرح التعرف الآلي على الأحداث مهمة مثيرة للاهتمام فيما يتعلق بتحديثات وسائل التواصل الاجتماعي. في حين يمكن الحصول على مجموعة كبيرة من التغريدات تكفي للكشف عن الاتجاهات والأحداث الدولية، تظل هناك مشكلة تطوير وتقييم خوارزميات قادرة على التعامل مع تحديثات بهذا الحجم.

لا تستخدم غالبية منهجيات التعرف على الأحداث الأنطولوجيات أو غيرها من مصادر المعلومات الدلالية. هناك فئة من الأساليب التي تطبق عملية التجميع على التغريدات [255-257] أو مشاركات المدونات [258]. على سبيل المثال، استخدمت دراسة [259] منهجية من هذا القبيل لكشف الزلازل في اليابان بناءً على أساس التغريدات التي تتضمن معلومات تحديد المواقع الجغرافية. وبالمثل، جرى التعامل مع الكلمات الفردية كإشارات موجية (wavelet signals) من أجل استكشاف تجمّعات مصطلحات ذات أهمية زمنية [260].

بمجرد الكشف عن حدث ما في تحديثات وسائل التواصل الاجتماعي، تصبح المشكلة التالية وهي كيفية إنتاج عناصر توصيف (descriptors) موضوعية مفيدة خاصة بهذا الحدث. جرى في الآونة الأخيرة الجمع بين المعلومات التبادلية النقطية (point-wise mutual information) والمعلومات الجغرافية والزمنية الخاصة بالمستخدم، وذلك من أجل الحصول على سلاسل ن-جرام (n-gram) لتوصيف الأحداث من التغريدات [261]. من خلال جعل الخوارزمية حساسة للموقع الأصلي، من الممكن رؤية ما يتداوله الناس في موقع معين بشأن حدث ما (كالأشخاص المتواجدين في الولايات المتحدة)، وكيف يختلف ذلك عن التغريدات الأخرى (كالأشخاص الموجودين في الهند).

يمكن الإشارة إلى مجموعات الأحداث الموجودة في تسلسل أكبر على أنها قصص ملاحم (sagas)، وقد تكون أحداثًا حقيقية تمامًا بحد ذاتها، أو قد تكون مكوناتها الفردية متناسقة بحد ذاتها. تشير دراسة [135] - التي اقتبست مثال من مؤتمر أكاديمي- إلى أن التغريدات قد تشير إلى المؤتمر ككل، أو إلى حدث فرعي محدد مثل العروض التي تجري في وقت ومكان معين. باستخدام المعلومات الدلالية الخاصة بالمؤتمر وأحداثه الفرعية من شبكة بيانات (Web of Data)، تتم مواءمة التغريدات مع تلك الأحداث الفرعية بصورة تلقائية، وذلك باستخدام أساليب التعلم الآلي. يشمل هذا الأسلوب مرحلة تعزيز المفهوم تُستخدم فيها أداة Zemanta لإضافة مفاهيم قاعدة البيانات DBpedia كشروحات إلى كل تغريدة. توصف التغريدات دلاليًّا باستخدام أنطولوجيا SIOC وأنطولوجيات الحضور الإلكتروني (Online Presence) (راجع القسم ٨-٢).

في الدراسة [245] جرى اقتراح أسلوب دلالي آخر يستند إلى الكيانات لكشف الأحداث الفرعية التي تستخدم معلومات أساسية جرى إعدادها يدويًّا عن الحدث (كأسماء الفرق واللاعبين في ألعاب الكريكت)، بالإضافة إلى معرفة ذات نطاق محدد مأخوذة من موقع ويكيبيديا (كالأحداث الفرعية المتعلقة بالكريكت كالخروج من اللعب). علاوة على إضافة هذه المعلومات الدلالية إلى التغريدات كشروحات، يستخدم هذا الأسلوب حجم التغريدات (مثلما هو الحال مع أسلوب [262]) وكذلك وتيرة نشر التغريدات المكررة كمؤشرات خاصة بالأحداث الفرعية. غير أن وجه

القصور في هذه المنهجية يأتي من الحاجة للقيام بتدخل يدوي، وهو ما لا يكون عمليًّا في العادة خارج عدد محدود من مجالات التطبيق.

٨-٣-٤ تمييز المشاعر وتعدين الآراء

يعدُّ وجود مواقع إلكترونية تحظى بالشعبية مكرسة للتقييمات وآراء المستخدمين حول المنتجات والخدمات بمنزلة إقرار بأهمية الدافع الموجود لدى الإنسان لنشر ما يشعر أو يفكر به على الإنترنت. وبالنظر لكون النوع الأكثر شيوعًا من رسائل تويتر متعلقًا بـ«الذات واللحظة» [263]، فمن المتوقع أن يتحدث المستخدمون عن مزاجهم وآرائهم. يجادل (بولين وآخرون) [194] بأن المستخدمين يعبرون عن مزاجهم الشخصي في تغريدات تتعلق بهم شخصيًّا وأيضًا في رسائل تتعلق بأشخاص آخرين. هناك دراسة أخرى [264] تقدر أن ١٩٪ من رسائل المدونات المصغّرة تذكر علامة تجارية معينة، فيما تحتوي ٢٠٪ من تلك الرسائل على المشاعر المتعلقة بتلك العلامة التجارية.

تحمل هذه الأفكار والآراء قيمة عظيمة. على سبيل المثال، يمكن أن تعكس عملية التحليل الجماعي لتلك الآراء صورة واضحة عن المزاج العام، وهو ما يتيح استكشاف ردود الأفعال على الأحداث العامة الجارية [194] أو ملحوظات على أفراد أو حكومات أو منتجات أو خدمات معينة [265]. يمكن استخدام المعلومات الناتجة لتحسين الخدمات أو صياغة السياسات العامة أو جني الأرباح من أسواق الأسهم.

تنطلق شرارة أنشطة المستخدمين على وسائل التواصل الاجتماعي في الغالب بفعل أحداث معينة وما يتصل بها من كيانات (كالأحداث الرياضية والاحتفالات والأزمات والمقالات الإخبارية والأشخاص والمواقع) وموضوعات (كالاحتباس الحراري والأزمات المالية وإنفلونزا الخنازير). من أجل تضمين هذه المعلومات، كانت هناك حاجة لوجود منهجيات واعية دلاليًّا واجتماعيًّا.

هناك العديد من التحديات الكامنة في تطبيق أساليب تعدين الآراء وتحليل المشاعر على وسائل التواصل الاجتماعي [266]. يمكن القول: إن المشاركات المصّغرة هي الأكثر صعوبة من بين أنواع النصوص المختلفة.

عندما يتعلق الأمر بتعدين الآراء، وذلك نظرًا لكونها لا تحتوي على الكثير من المعلومات السياقية وتفترض الكثير من المعرفة الضمنية. يعدُّ الغموض مشكلة خاصة؛ لأنه لا يمكننا الاستفادة بسهولة من معلومات الإشارة المشتركة (coreference). فخلافًا لمشاركات وتعليقات المدونات، لا يجري ترتيب التغريدات في العادة لتندرج تحت موضوعات محادثات، وتظهر بصورة منفصلة جدًّا عن التغريدات الأخرى. تتسم التغريدات أيضًا بتباين لغوي أكبر وتميل إلى أن تكون أقل تقيّدًا بالقواعد النحوية مقارنة بالمشاركات الطويلة، كما تحتوي على قواعد غير تقليدية لكتابة الأحرف الكبيرة، ويتكرر فيها استخدام رموز التعبيرات والاختصارات وعلامات الهاشتاغ، وهو ما يمكن أن يشكل جزءًا مهمًّا من المعنى. في العادة، تحتوي التغريدات أيضًا على استخدام كبير للسخرية والتهكم، وهما من الأشياء التي يصعب على الآلات اكتشافها على وجه التحديد. من جهة أخرى، يمكن أن تكون طبيعتها الموجزة مفيدة من ناحية تركيزها على الموضوعات بصورة أكثر صراحة، فنادرًا جدًّا ما تكون تغريدة واحدة متعلقة بأكثر من موضوع واحد، مما يساعد في إزالة الغموض عن طريق التأكيد على الصلة الظرفية.

خلافًا لبعض أدوات تحليل المشاعر على المستوى المفاهيمي المصممة حديثًا لتحليل النصوص، كتقييمات المنتجات والرحلات (كما ناقشنا في القسم ٧-٦) التي تركز على المنهجيات المعتمدة على الخصائص، تستخدم غالبية أساليب تعدين المشاعر والآراء التي جرى اختبارها على وسائل التواصل الاجتماعي قدرًا ضئيلاً أو معدومًا من الدلالات. على سبيل المثال، تصنف دراسة [267، 268] التغريدات إلى تغريدات تحتوي على مشاعر إيجابية أو سلبية أو محايدة، وذلك بناءً على سلاسل ن-جرام (n-grams) والمعلومات المتعلقة بأقسام الكلام، في حين تستخدم دراسة [269] معجمًا دلاليًّا لإضافة الشروح إلى المشاعر الإيجابية والسلبية بشكل مبدئي في التغريدات ذات الصلة بالأحداث السياسية.

يؤدي استخدام هذا النوع من المعلومات إلى بروز مشكلة تبعثر البيانات. تبيّن دراسة (سيف وآخرون) [133] أن دقة تصنيف القطبية تتحسن باستخدام المفاهيم الدلالية، بدلاً من كلمات من قبيل آيفون. تستخدم هذه المنهجية برنامج AlchemyAPI لإضافة الشروح الدلالية إلى ٣٠ فئة من فئات الكيانات، ومن أكثرها شيوعًا فئات كالأشخاص

(Person) والشركات (Company) والمدن (City) والدول (Country) والمؤسسات (Organization). يجري تقييم هذا الأسلوب بواسطة قاعدة بيانات ستانفورد لمشاعر التغريدات[1]، وقد ثبت أن أداءها يتفوق على الأساليب العصرية الخالية من الدلالات، بما في ذلك أسلوب [268].

استُخدمت عملية إضافة الشروح الدلالية لغرض إجراء مهام تعدين الآراء الأكثر صعوبة. على وجه الخصوص، تحدد دراسة [270] هوية الأشخاص والأحزاب السياسية والبيانات التي تعرب عن رأي ما في التغريدات باستخدام أداة للتعرف على الكيانات استنادًا إلى القواعد، بالإضافة إلى معجم «affect» الذي يضم مجموعة من الكلمات ذات الصلة بالمشاعر المأخوذة عن قاعدة بيانات WordNet. يستخدم التحليل الدلالي الذي يجري بعد ذلك أنماطًا لتوليد ثلاثيات تمثل أصحاب الآراء ونِيّات المصوتين. يجري التعامل مع النفي (Negation) من خلال جمع وتسجيل الأنماط البسيطة من قبيل «ليس مفيدًا» أو «ليس مثيرًا» واستخدام تلك الأنماط لنفي أحكام المشاعر المستخرجة. جرى توسيع نطاق هذا العمل في وقت لاحق عن طريق إضافة الدلالات إلى المصطلحات السياسية (المرتبة حسب تسلسل هرمي) وأعضاء البرلمان في أداة لتحليل النقاشات التي دارت في تويتر حول الانتخابات البريطانية في عام ٢٠١٥ [271].

٨-٣-٥ الربط بين الوسائط الإعلامية

إضافة إلى كونها وثيقة الصلة بالأحداث الدائرة في العالم الحقيقي، تعني الطبيعة الموجزة لرسائل تويتر وفيسبوك أنه لا يمكن فهم المشاركات القصيرة في الغالب من دون الرجوع إلى سياق خارجي. وفي حين تحتوي بعض المشاركات فعليًّا على عنوانات URL، إلا أن غالبيتها لا تحتوي على تلك الروابط. لذا تكون هناك حاجة لاستخدام أساليب لربط الوسائط المختلفة بعضها بعض وإثرائها بالسياق والدلالات بصورة تلقائية.

1- http://alt.qcri.org/semeval2017/task8/

تربط دراسة (أبيل وآخرون) [134] التغريدات بالتقارير الإخبارية من أجل تحسين دقة عملية إضافة الشروح الدلالية إلى التغريدات. في هذه الدراسة، يجري البحث في عدد من استراتيجيات ربط التغريدات بالوسائط، مثل الاستفادة من عنوانات URL الموجودة في التغريدة، وشبه قيمة TF-IDF (تكرار المصطلح/ عكس تكرار المستند) بين التغريدة والمقالة الإخبارية وعلامات هاشتاغ وأوجه الشبه المستندة إلى الكيانات (يجري التعرف على الكيانات والموضوعات الدلالية بواسطة خدمة OpenCalais)، حيث تكون أوجه الشبه المستندة إلى الكيانات الأفضل للتغريدات التي لا تتضمن عنوانات URL. هذه المنهجية شبيهة باستراتيجية الربط المستندة إلى العبارات المفتاحية لمطابقة لقطات الفيديو الإخبارية مع الصفحات الإخبارية الإلكترونية [272]. تذهب دراسة [273] خطوة أبعد من ذلك، وذلك من خلال جمع محتوى وسائل التواصل الاجتماعي حول التغير المناخي من تويتر ويوتيوب وفيسبوك مع الأخبار على الإنترنت، على الرغم من أن تفاصيل الخوارزمية المستخدمة للربط بين الوسائط المختلفة لم تقدم في هذه الورقة البحثية.

توصلت دراسة متعمقة سعت للمقارنة بين أخبار تويتر وجريدة نيو يورك تايمز [274] إلى ثلاثة أنواع من الموضوعات، وهي الموضوعات المستندة إلى الأحداث، والموضوعات المستندة إلى الكيانات، والموضوعات طويلة الأمد. كما يجري تصنيف الموضوعات أيضًا إلى فئات مختلفة، بناءً على مجال الموضوع. من بين الفئات التصنيفية، هناك تسع فئات مأخوذة من الفئات المستخدمة في جريدة النيويورك تايمز (كالفن والعالم والأعمال) بالإضافة إلى فئتين خاصتين بتويتر (الأسرة والحياة، وتويتر). تعدُّ فئة الأسرة والحياة الفئة السائدة في تويتر (تسمى فئة «أنا الآن» في دراسة [263])، سواءٌ من حيث عدد التغريدات وعدد المستخدمين. أظهرت المقارنة الآلية المستندة إلى الموضوعات أن التغريدات تعج بالموضوعات المستندة إلى الكيانات، وتقل تغطية هذا النوع من الموضوعات كثيرًا عن غيره من أنواع الموضوعات في وسائل الإعلام التقليدية.

لتجاوز نطاق الأخبار والتغريدات، هناك حاجة لإجراء بحوث في المستقبل حول مسألة الربط بين الوسائط المختلفة. على سبيل المثال، يقوم بعض المستخدمين بنقل

تغريداتهم إلى حساباتهم على فيسبوك، وهناك يستقطب محتوى تغريداتهم تعليقات المستخدمين بصورة منفصلة عن أي ردود تتم على التغريدات الأصلية أو أي إعادة نشر لها من قبل المستخدمين الآخرين. وبالمثل، يمكن الجمع بين التعليقات الموجودة على صفحة مدونة ما والتغريدات التي تتناول تلك الصفحة، وذلك من أجل تكوين رؤية أكثر شمولية.

٨-٣-٦ تحليل الشائعات

هناك نوع محدد من أنواع التحليل الدلالي لوسائل التواصل الاجتماعي، وهو تحليل الشائعات. أظهرت الأبحاث في البداية الضرر الذي يمكن أن يلحقه نشر الشائعات المزيفة على المجتمع، وكذلك الانتشار البطيء للتغريدات التي تكشف زيف تلك الشائعات [275، 276]. لذا فإن القدرة على تحديد دقة المعلومات المنشورة على وسائل التواصل الاجتماعي تعدُّ مهمة. غير أن عملية التأكد من صحة الشائعات عادة ما تكون صعبة [390]، وذلك لأنه لا بد من جمع أكبر عدد ممكن من الآراء والشهادات ومعاينتها من أجل التوصل إلى حكم نهائي. تشمل أمثلة الشائعات التي جرى إثبات عدم صحتها لاحقًا، بعد تداولها على نطاق واسع في البداية، هزة أرضية وقعت في عام 2010 في دولة تشيلي، حيث انتشرت شائعات حول انفجار بركان وصدور تحذيرات عن موجات تسونامي في مدينة فالبارايسو على موقع تويتر [277]. من الأمثلة الأخرى أعمال الشغب التي حدثت في إنجلترا في عام 2011، حيث زعمت شائعات كاذبة أن مثيري الشغب كانوا ينوون مهاجمة مستشفى برمنغهام للأطفال وأن الحيوانات قد هربت من حديقة لندن للحيوانات [278].

تتمثل الخطوة الأولى لتحليل الشائعات في اكتشاف التغريدات المتعلقة بالشائعات [279، 280].

من الأعمال المؤثرة الدراسة التي أجراها (ميندوزا وآخرون) [277]، حيث قاموا بإجراء تحليل يدوي لـ٧ حقائق مؤكدة و٧ شائعات كاذبة حول الزلزال الذي وقع في تشيلي في عام ٢٠١٠، علمًا أن كل شائعة تضمنت نحو ١,٠٠٠ تغريدة. بعد ذلك جرى تصنيف التغريدات يدويًّا حسب موقفها تجاه الشائعة، سواءٌ أكان موقفها مؤكدًا

أم نافيًا أم مشككًا أم غير معروف أم غير ذي صلة. أظهرت الدراسة أنه قد اتضح إنكار نسبة أعلى بكثير من التغريدات المتعلقة بالشائعات الكاذبة للشائعة المعنية (٥٠٪ تقريبًا)، وهو ما يناقض الشائعات التي اتضحت صحتها لاحقًا، حيث لم تتجاوز نسبة التغريدات النافية للشائعة ٠,٣٪ فقط. بناءً على ذلك، ادعى القائمون على الدراسة أنه يمكن الكشف عن الشائعات باستخدام التحليل الجمعي للمواقف التي تعكسها التغريدات.

شجع هذا الأمر على إجراء مجموعة ضخمة من الأبحاث في وقت لاحق حول تصنيف مواقف الشائعات. من بين المنهجيات الأولى منهجية دراسة (قَزوِينيان وآخرون) [281] التي صنّفت كل تغريدة بصورة آلية على أنها إما تغريدة داعمة أو نافية أو مشككة لشائعة معينة. غير أنهم قرروا الدمج بين التغريدات النافية والتغريدات المشككة وإدراجها تحت فئة واحدة، محولين العملية إلى إشكالية تصنيف ثنائي تنقسم إلى قسم داعم مقابل قسم نافٍ أو مشكك. تستخدم دراسة حميديان ودياب [282] متجهات التغريدات الكامنة (Tweet Latent Vectors) لتقييم قدرة عملية التصنيف الثنائي لمواقف التغريدات إلى مواقف داعمة أو نافية لشائعة ما. كما تشير الدراسة إلى أي مدى يمكن استخدام نموذج مدرّب على تغريدات تاريخية لتصنيف تغريدات جديدة حول الشائعة نفسها.

أرجعت أعمال بحثية جرت في الآونة الأخيرة هذا التصنيف إلى التصنيف الثلاثي الأكثر واقعية [283]. تشمل المنهجيات البارزة الأخرى منهجية (ليو وآخرون) [284] الذين استحدثوا أساليب تعتمد على القواعد لتصنيف مواقف التغريدات، ويتفوق أداء هذه الأساليب على أداء [281]. وبالمثل، تستخدم دراسة [279] التعبيرات النمطية (regular expressions) لتصنيف مواقف الشائعات.

في جميع تلك الحالات، يتمثل التحدي الأكبر في تعميم المنهجية المتبعة لتشمل الشائعات الجديدة التي لم تظهر من قبل والتي تختلف عادة عن التغريدات التي يصادفها برنامج التصنيف في بيانات التدريب. تجاهلت الأعمال السابقة التمييز بين الشائعات القديمة والجديدة وجمعت بين التغريدات المتعلقة بجميع الشائعات باستخدام أسلوب التصديق التبادلي (cross-validation). تحدد دراسة أجريت حديثًا وتناولت تصنيف

مواقف التغريدات تجاه الشائعات [285] المشكلة على أنها عبارة عن انتقال أثر التعلم (transfer learning)، وقيّمت الشائعات التي لم تظهر فقط. تناولت دراسة (زينج وآخرون) [286] استخدام ثلاثة مُصنّفات (Random Forest وNaive Bayes و Logistic Regression) لتصنيف المواقف تجاه الشائعات بصورة آلية على الشائعات المخفية، لكنها ركزت فقط على تعريف المشكلة بثنائية الدعم/ النفي.

يتمثل التحدي الأساسي أمام الباحثين في مجال شائعات ووسائل التواصل الاجتماعي في عدم وجود قاعدة بيانات ضخمة ومتوفرة على نطاق واسع. يهدف تحدي 2017 RumourEval إلى التعامل مع هذه المشكلة[1]، بالإضافة إلى توفير آلية للمقارنة بين الوسائل المختلفة الخاصة بالتحقق من صحة الشائعات وتصنيف مواقف الشائعات. من بين مجموعات البيانات التي ظهرت مؤخرًا مجموعة بيانات [287].

٨-٣-٧ النقاش

على الرغم من تحقيق بعض الاختراقات بصورة فعلية، إلا أن الأساليب الحالية المستخدمة لإضافة الشروح الدلالية إلى تحديثات وسائل التواصل الاجتماعي تحمل الكثير من أوجه القصور. في البداية، تتعامل غالبية الأساليب مع المشكلات السطحية المتمثلة في استخراج الكلمات المفتاحية والموضوعات، في حين لا تحقق أساليب تمييز الكيانات والأحداث المبنية على الأنطولوجيات نتائج ذات دقة وقدرة على الاسترجاع أعلى بكثير من النتائج التي يجري الحصول عليها عند التعامل مع الوثائق ذات النصوص الطويلة. من بين الطرق المتبعة لتحسين الأداء الآلي السيئ في الوقت الحالي أسلوب التعهيد الجماعي (crowdsourcing). على سبيل المثال، يجمع نظام ZenCrowd [288] بين خوارزميات مستخدمة لربط الكيانات بالمُدخلات البشرية على نطاق واسع عبر نظام المهام المتناهية الصغر عبر خدمة Amazon Mechanical Turk لإنجاز المهام. بهذه الطريقة، لا يتم إظهار الإشارات النصية (textual mentions) التي يمكن ربطها آليًّا وبمستوى ثقة مرتفع بالحالات (instances) الموجودة في سحابة البيانات المفتوحة المترابطة (LOD Cloud) لمضيفي الشروح الدلالية من البشر. لا تجري استشارة الحالات

1- http://ln.ontotext.com/KIM

(instances) إلا عندما يكون حلها صعبًا، وهو ما لا يؤدي إلى تحسين جودة النتائج فحسب، بل يحد أيضًا من كمية التدخلات اليدوية المطلوبة. سوف نعود إلى تناول موضوع التعهيد الجماعي (crowdsourcing) بمزيد من التفصيل في القسم ١٠-٢.

هناك طريقة أخرى لتحسين عملية إضافة الشروح الدلالية إلى محتوى وسائل التواصل الاجتماعي، وهي استخدام المعرفة الضخمة المتوفرة على شبكة البيانات (Web of Data) استخدامًا أفضل. في الوقت الحالي، تقتصر تلك المعرفة على ويكيبيديا والمصادر المشتقة منها (كقاعدة بيانات DBpedia وYAGO). من التحديات الموجودة هنا تحدي الغموض. على سبيل المثال، تكون عناوين الأغاني والألبومات في MusicBrainz شديدة الغموض، كما تتضمن كلمات شائعة (مثل أمس) وكلمات التوقف (The, If) [244]. بناءً على ذلك، قد تكون هناك حاجة لإجراء خطوة تصنيف آلي للنطاق (domain)، وذلك لضمان استخدام مصادر البيانات المفتوحة المترابطة (LOD) ذات النطاق المحدد، مثل MusicBrainz، من أجل إضافة الشروح الدلالية إلى محتوى وسائل التواصل الاجتماعي التي تنتمي إلى النطاق المطابق فقط. من بين التحديات الأخرى تحدي الفاعلية والقابلية للتوسيع. في البداية، لا بد من أن تكون خوارزميات إضافة الشروح الدلالية فعالة في تعاملها مع اللغة المشوشة وغير المنظمة من حيث التركيبة النحوية التي تُستخدم في وسائل التواصل الاجتماعي. ثانيًا، بالنظر إلى حجم شبكة البيانات، فإن مهمة تصميم خوارزميات تستند إلى الأنطولوجيات وقادرة على تشغيل قواعد المعرفة الضخمة هذه واستعلام البيانات منها، مع الحفاظ في الوقت ذاته على مستويات عالية من الإنتاجية الحسابية، ليست مهمة بسيطة.

تكمن العقبة الأخيرة أمام استخدام موارد شبكة البيانات (Web of Data) في كون المعلومات المعجمية المتاحة محدودة إلى حد بعيد. باستثناء الموارد المستندة إلى ويكيبيديا، فإن المعلومات المعجمية في باقي الموارد محدودة في الغالب ببطاقات RDF، وهو ما يحد بدوره من فائدتها باعتبارها مصدرًا للمعرفة لعمليات استخراج المعلومات وإضافة الشروح الدلالية المستندة إلى الأنطولوجيات. ركزت إحدى مسارات الأبحاث التي أجريت في الآونة الأخيرة على استخدام مصادر الويكاموس (Wiktionary) [دمج كلمتي ويكي وقاموس] [289] وهي مصادر معجمية متعددة اللغات ومبنية بصورة

تعاونية، وتعدُّ ذات أهمية خاصة لتحليل المحتوى المقدم من قبل المستخدم، وذلك نظرًا لكونها تحتوي على الكثير من التعابير الجديدة ويجري تحديثها بصورة متواصلة من قبل المساهمين. في اللغتين الإنجليزية والألمانية بالتحديد، يوجد أيضًا أعمال مستمرة حول إنشاء مصدر UBY [290] - وهو مصدر معجمي - دلالي واسع النطاق يعتمد على ويكيبيديا وقاعدة البيانات Wordnet، ولذا يعتمد بصورة غير مباشرة على مصادر البيانات المفتوحة المترابطة (LOD) الأخرى كذلك. هناك مسار مهم آخر وهو الأعمال التي تتعلق بالأنطولوجيات المستندة إلى اللغات [291]، التي اقترحت نموذجًا أكثر تعبيرًا لربط المعلومات اللغوية بعناصر الأنطولوجيات. وفي حين تعدُّ تلك الجهود خطوات في الاتجاه الصحيح، ما زالت هناك حاجة للقيام بالمزيد من العمل، ولا سيما بخصوص بناء أنظمة متعددة اللغات لإضافة الشروح الدلالية.

علاوة على ذلك، من البديهي أن تكون جودة أساليب إضافة الشروح الدلالية مرهونة ببيانات التدريب والتقييم الخاصة بها. تعدُّ عملية تدريب الخوارزميات على مجموعات بيانات وسائل التواصل الاجتماعي ذات المعيار الذهبي محدودة للغاية في الوقت الراهن. على سبيل المثال، يقل عدد التغريدات التي أضيفت إليها أنواع وأحداث كيانات الأسماء عن ١٠,٠٠٠ تغريدة في الوقت الراهن. لذلك توجد هناك حاجة ماسة لمكانز تقييم مشتركة وأكبر حجمًا ومكونة من شتى أنواع محتوى وسائل التواصل الاجتماعي. تعدُّ عملية إنشاء هذه المكانز عبر المنهجيات اليدوية التقليدية لإضافة الشروح الدلالية إلى النصوص باهظة الثمن، إن كان الهدف إنشاء عدد كبير من المكانز. ظلت الأبحاث التي تتناول المعايير الذهبية لعملية تقييم التمويل الجماعي محدودة، مع التركيز بصورة رئيسة على خدمة Amazon Mechanical Turk للحصول على مجموعات بيانات صغيرة (كالتغريدات ذات أنواع كيانات الأسماء) [292]. سوف نعود إلى هذا التحدي مرة أخرى في القسم ١٠-٢.

في مجال تحليل المشاعر، تناول الباحثون مشكلات اكتشاف قطبية المشاعر وتصنيف الموضوعية والتوقع عبر وسائل التواصل الاجتماعي وتنميط المزاج، غير أن غالبية الأساليب تستخدم قدرًا ضئيلاً أو معدومًا من الدلالات. إضافة إلى ذلك، يتسم تقييم تعدين الآراء بالصعوبة على وجه التحديد لعدد من الأسباب المنهجية (بالإضافة إلى

انعدام مصادر التقييم المشتركة التي سبقت مناقشتها أعلاه). أولاً، عادة ما تكون الآراء غير موضوعية، وليس من الواضح دائمًا مقصِد المؤلف. على سبيل المثال، لا يمكن للشخص أن يحدد ما إذا كان تعليق من قبيل «أحب المرأة اللطيفة فلانة»، عند غياب سياق إضافي، يعبر عن مشاعر إيجابية صادقة أو أنه يُستخدم على سبيل السخرية. لذا يميل الاتفاق بين مضيفي الشروح في البيانات التي تُضاف إليها الشروح يدويًّا إلى أن يكون متدنيًا، وهو ما يؤثر في موثوقية أي بيانات ذات معيار ذهبي يجري إنتاجها.

أخيرًا، تطرح تحديثات وسائل التواصل الاجتماعي عددًا من التحديات الإضافية العالقة حول أساليب تعدين الآراء والمشاعر:

الصلة: في وسائل التواصل الاجتماعي، يمكن أن تتشعب النقاشات والتعليقات بسرعة إلى موضوعات لا تمت بصلة للموضوع الأصلي، خلافًا لتقييمات المنتجات التي نادرًا ما تحيد عن الموضوع قيد النقاش.

تحديد الهدف: غالبًا ما يمكن أن يكون هنا عدم تطابق بين موضوع المشاركة المنشورة على إحدى وسائل التواصل الاجتماعي، الذي قد لا يكون بالضرورة موضوع المشاعر التي تحملها التغريدة. على سبيل المثال، في اليوم التالي لوفاة ويتني هيوستون، عرض موقع TwitterSentiment والمواقع المشابهة أن الغالبية العظمى من التغريدات المتعلقة بويتني هيوستون كانت سلبية، لكن جميع تلك التغريدات تقريبًا كانت سلبية فقط لأن الناس كانوا يشعرون بالحزن على وفاتها، وليس لأنهم كانوا يكرهونها.

التقلب بمرور الوقت: بشكل أكثر تحديدًا، يمكن أن تتغير الأفكار بصورة درامية بمرور الوقت، من كونها أفكارًا إيجابية إلى أفكار سلبية والعكس. للتعامل مع هذه المشكلة، يمكن ربط الأنواع المختلفة للآراء الممكنة باعتبارها خصائص أنطولوجيا بالأنواع التي تصف الكيانات والحقائق والأحداث المكتشفة عبر أساليب إضافة الشروح الدلالية، وهي شبيهة بتلك الموجودة في [293] التي تهدف إلى التحكم في تطور الكيانات بمرور الوقت. يمكن توثيق الآراء والمشاعر المستخرجة زمنيًّا ومن ثمّ تخزينها في قاعدة معرفة يتم تعزيزها باستمرار مع إضافة محتوى وآراء جديدة. هناك إشكالية متعلقة بهذا الموضوع، وهي كيف يمكن اكتشاف الآراء المستجدة، بدلاً من إضافة المعلومات الجديدة إلى رأي موجود مسبقًا للكيان المعني. أيضًا هناك حاجة لتدوين

التناقضات والتغييرات واستخدامها لمراقبة الاتجاهات المستجدة بمرور الوقت، ولا سيما عبر تجميع الآراء.

تجميع الآراء: هناك تحدٍّ آخر وهو نوع التجميع الذي يمكن تطبيقه على الآراء. في مهمة إضافة الشروح الدلالية المستندة إلى الكيانات، يمكن تطبيق ذلك على المعلومات المستخرجة بطريقة سهلة ومباشرة، إذ يمكن دمج البيانات معًا إذا لم يوجد أي تباينات فيما بينها، على سبيل المثال، فيما يتعلق بخصائص كيان من الكيانات. لكن سلوك الآراء يختلف هنا، حيث يمكن إرفاق عدة آراء بكيان واحد وينبغي نمذجتها بصورة منفصلة، ونحن نؤيد تعبئة قاعدة معرفة لهذا الغرض. هناك سؤال مهم يتعلق بما إذا كان ينبغي على الباحث تخزين متوسط الآراء المكتشفة ضمن حيّز زمني محدد (مثلما تفعل الأساليب المستخدمة حاليًّا لعرض الآراء في صيغة مرئية)، أو ما إذا كان يُفضل استخدام منهجية أكثر تفصيلاً، مثل نمذجة المصادر وقوة الآراء المتضادة وطبيعة التغير الذي يطرأ عليها بمرور الوقت. هناك سؤال مهم آخر في هذا السياق، ويتعلق بإيجاد تجمعات الآراء التي يجري التعبير عنها على وسائل التواصل الاجتماعي وفقًا للمجموعات والشرائح الديموغرافية والأوساط الجغرافية والاجتماعية المؤثرة.

وعلى هذا النحو، تتطلب الطبيعة الاجتماعية المعتمدة على الرسوم البيانية للتفاعلات استخدام أساليب جديدة لتجميع الآراء.

الفصل التاسع
التطبيقات

توجد العديد من التطبيقات المتنوعة في مجال إضافة الشروح والتعليقات الدلالية، ومنها البحث الدلالي، وهو إيجاد المستندات التي يرد فيها ذكر مفهوم/ حالة واحدة أو أكثر داخل أنطولوجيا أو بيانات مفتوحة مترابطة، وبناء نماذج المستخدم الاجتماعية الدلالية، بما فيها البيانات الديموغرافية واهتمامات المستخدمين والسلوك الإلكتروني ونمذجة المجتمعات الإلكترونية والتجسيد البصري للمعلومات بالاستناد إلى الدلالات. تستغل كل هذه التطبيقات مُخرجات المراحل السابقة في عملية معالجة النص، ومنها تمييز كيانات الأسماء وربطها واستخراج العلاقات والمصطلحات وتحليل المشاعر وغيرها.

يُقدم هذا الفصل كل تطبيق من هذه التطبيقات على حدة، ولن يقتصر الشرح على المبادئ الأساسية لكل تطبيق من هذه التطبيقات، بل سيشير أيضًا إلى عددٍ من الأمثلة الأساسية المأخوذة من الأدبيات. ثم نختتم الفصل بنقاش للأسئلة المطروحة والاتجاهات المستقبلية.

٩-١ البحث الدلالي

يعد طرح مقدمة ومراجعة متعمقة للأدبيات الراهنة في مجال البحث الدلالي خارج نطاق هذا الكتاب، لكن يُنصح القارئ بمراجعة [294، 295] لمزيد من التفاصيل. ستقتصر المادة المقدمة في هذه الفقرة على لمحة عامة فقط.

يعدُّ البحث الدلالي داخل الوثائق مهمة تُعنى بإيجاد معلومات ليس بناءً على مدى توفر كلمات معينة فحسب، بل أيضًا بناءً على معنى هذه الكلمات [296، 297]. هذه المهمة هي صيغة معدلة من مهمة استرجاع المعلومات (IR) التقليدية، لكنها تختلف في أنه يجري استرجاع المستندات بناءً على مدى صلتها بالمفاهيم الواردة داخل الأنطولوجيا، بالإضافة إلى الكلمات. غير أن الفرضية الأساسية في كلتا المهمتين متطابقة إلى حد بعيد، فما يحدد سمات مستند معين هي مجموعة بطاقات التصنيف التي تشكل محتوى الوثيقة، بصرف النظر عن هيكلها. وفي حين تعدُّ منهجية استرجاع المعلومات الأساسية أن جذور الكلمات هي بطاقات تصنيف، هناك جهودٌ كبيرة بُذلت

من أجل استخدام معاني الكلمات أو المفاهيم المعجمية [298، 299] في عملية الفهرسة والاسترجاع. في حالة البحث الدلالي، ما تتم فهرسته عادة يكون مجموعة من الكلمات ومفاهيم أنطولوجيا توصل معنى قسم من هذه الكلمات (مثال: كامبريدج هو موقع)، وهناك خيار تحديد معنى العلاقات القائمة بين هذه المفاهيم (مثال: كامبريدج توجد في المملكة المتحدة) [296]. يتيح المثال الثاني لشخص ما يبحث عن مستندات متعلقة بالمملكة المتحدة العثور أيضًا على وثائق تذكر كامبريدج.

غير أن كلمة كامبريدج (وكذلك العديد من الأسماء والكلمات الأخرى) لها معانٍ عدة، أي أنها غامضة. فقد تشير كلمة «كامبريدج» إلى مدينة كامبريدج في المملكة المتحدة أو مدينة كامبريدج في ولاية ماساتشوستس الأمريكية أو جامعة كامبريدج ...الخ. وبالمثل، قد تحمل البطاقات التصنيفية المختلفة المعنى نفسه، مثال، نيويورك و«بيج أبل» (التفاحة الكبيرة). لذا يحاول البحث الدلالي تقديم نتائج أكثر دقة وصلة للمستخدمين، وذلك باستخدام التعليقات والشروحات الدلالية والمعرفة الخارجية المُشفّرة عادة في الأنطولوجيات و/ أو مصادر البيانات المفتوحة المترابطة.

من الناحية العملية، تُستخدم مُخرجات أساليب إضافة الشروحات الدلالية (كالتي ورد نقاشها في الفصل الخامس) لتمكين المستخدمين من إيجاد وثائق تذكر حالة (instance) وفئة (class) و/ أو علاقة (relation) واحدة أو أكثر. تدعم بعض منصات البحث الدلالي الاستعلامات التي تخلط بين الكلمات المفتاحية التي تكون على شكل نص حر والشروحات الدلالية بل وحتى استعلامات لغة «سباركل» (SPARQL). تقدم معظم أدوات استرجاع المعلومات أيضًا خاصية تصفح الوثائق، وكذلك قدرات تنقيح نتائج البحث. وبسبب إمكانية وجود مئات من التعليقات الدلالية في الوثائق (ولا سيما في حال وجود تعليقات دلالية مصاحبة لكل مفهوم يرد ذكره في الوثيقة)، فإن عملية استرجاع الشروح الدلالية في مجموعات كبيرة من الوثائق هي عملية شديدة الصعوبة.

تختلف عمليات البحث المستندة إلى الشروح عن عمليات استرجاع المعلومات التقليدية، وذلك بسبب التمثيل الرسومي الكامن فيها الذي يؤدي إلى تشفير المعلومات

المهيكلة عن نطاقات النصوص داخل الوثيقة. تختلف المعلومات المشفّرة عن الكلمات ونماذج الربط بين الوثائق المستخدمة من طرف جوجل وغيرها من محركات البحث. كما تشير العديد من الشروح الدلالية إلى الأنطولوجيات بواسطة معرفات الموارد الموحدة (URIs). وفي حين قد تساعد فهارس النصوص الكاملة (full-text) المعززة في رفع كفاءة عملية الوصول، إلا أن متطلبات تخزين البيانات قد تكون ضخمة جدًّا، وذلك مع تنامي عدد العناصر في مجموعات الشروح الدلالية. لذلك جرى البحث عن حلول مختلفة ذات كفاءة عليا.

يكمن وجه الاختلاف الرئيس عن محركات البحث الخاصة بالويب الدلالي، مثل محرك Swoogle [300] في أن التركيز يكون على عملية إضافة التعليقات، ومن ثم استخدامها في عملية إيجاد الوثائق، بدلاً من الاستعلام داخل الأنطولوجيات أو تصفح هياكل الأنطولوجيات. وبالمثل، تميل واجهات البحث والتصفح متعدد الأوجه المستند إلى الدلالات، مثل facet/ [301]، إلى أن تكون مستندة إلى الأنطولوجيات، بينما تميل واجهات البحث والتصفح متعدد الأوجه المستند إلى الشروحات (راجع KIM أدناه) إلى إخفاء الأنطولوجيا ومحاكاة عمليات البحث «التقليدية» متعددة الأوجه المستندة إلى سلاسل الكلمات.

٩-١-١ ما البحث الدلالي؟

لفهم الأنواع المختلفة من مهام ومنهجيات البحث الدلالي، من المفيد أن نضع في الاعتبار جانبين، وهما: (أ) ما يجري البحث عنه و(ب) ما النتائج. سوف نناقش هذين الأمرين واحدًا تلو الآخر.

بخصوص الشيء الذي يجري البحث عنه، هناك ثلاثة أنواع رئيسة من المحتوى التي ينبغي أخذها بعين الاعتبار:

الوثائق: هذا النوع من البحث هو بحث النص الكامل التقليدي، حيث تأتي الردود على الاستعلامات بناءً على التوارد المشترك للكلمات في محتوى النص. على سبيل المثال، تكون نتيجة استعلام مثل «جامعة كامبريدج» جميع المستندات التي تحتوي على كلمتي

كامبريدج و/ أو جامعة في مكان ما. لا يعني ذلك أن النتائج هي مستندات تتعلق بتلك الجامعة فقط. هذا النوع من البحث تواجهه مشكلات بخصوص الإجابة على الاستعلامات التي تكون من نوع الكيانات، على سبيل المثال، ما المدن البريطانية التي يكون عدد سكانها أقل من ١٠٠,٠٠٠ نسمة.

الأنطولوجيات والمعارف الدلالية الأخرى مثل LOD: هذا البحث هو بحث داخل بيانات مهيكلة رسمية، يجري التعبير عنها بـRSD [302] أو OWL [303]، وتُخزن في قاعدة بيانات أو مستودع دلالي. ونتيجة لذلك، يجري التعبير عن مثل هذا النوع من الاستعلامات الرسمية بواسطة لغات استعلام مهيكلة مثل لغة «سباركل» (SPARQL) [304] أو لغة الاستعلامات البنيوية (SQL). غالبًا ما يُشار إلى هذا النوع من البحث بالبحث الدلالي، وذلك لكونه يستخدم الدلالات وأسلوب الاستنباط لإيجاد المعرفة الرسمية (formal knowledge) المطابقة. في هذا الفصل، سوف نشير إلى هذا النوع من البحث باسم البحث المستند إلى الأنطولوجيا. يناسب هذا النوع من البحث بصفة خاصة الرد على الاستعلامات التي تكون من نوع الكيانات كالمثال الذي أوردناه أعلاه.

المستندات والمعرفة الرسمية كليهما: هذا هو ما يشير إليه هذا الفصل بالبحث الدلالي في المستندات، أو البحث متعدد النماذج [297] أو بحث النص الكامل الدلالي [305]. يعتمد هذا النوع من البحث على محتوى المستندات والمعرفة الدلالية، وذلك من أجل الإجابة على استعلامات من قبيل: «فيضانات في مدن في المملكة المتحدة» أو «فيضانات في مناطق تبعد ٥٠ ميلاً عن شيفيلد.» في هذه الحالة، تكون المعلومات المتعلقة بالمدن الموجودة في المملكة المتحدة أو التي تقع على بعد ٥٠ ميلاً عن شيفيلد ناتجة عن عملية بحث مستندة إلى أنطولوجيا. بعبارة أخرى، يجري البحث هنا داخل محتوى المستند ويكون البحث عن الكلمات المفتاحية ومؤشر الكيانات التي تتضمن شروحًا دلالية موجودة داخل هذه المستندات، وكذلك المعرفة الرسمية.

وفيما يتعلق بالنتائج التي تنتج عن عمليات البحث، هناك أربعة أنواع رئيسة هي:

المستندات: تعطي عملية البحث قائمة مصنفة من المستندات، وعادة ما تُعرض هذه المستندات بعنواناتها، مع إمكانية عرض بعض البيانات الوصفية (مثال: المؤلف). هذا النوع من البحث عادة ما ينتج عن عمليات بحث النص الكامل، على الرغم من أن بعضها يتضمن مقتطفات أيضًا.

المستندات + مقتطفات تبرز أهم النتائج: بالإضافة إلى عنوانات المستندات، تعطي عملية البحث مجموعة واحدة أو أكثر من المقتطفات، مع إبراز النتائج التي تتطابق مع الاستعلام، وذلك في محاولة للتوضيح للمستخدم السبب وراء كون هذه الوثيقة ذات صلة باستعلامه. في العادة تقوم أنظمة البحث الدلالي بعرض المستندات المتطابقة مع الاستعلام بهذه الطريقة، ومن الأمثلة على تلك الأنظمة نظام KIM [296] ونظام Mímir [297] ونظام Broccoli [306].

تلخيص المعلومات: هذه العملية هي عبارة عن عرض المعرفة الرسمية في صيغة يمكن للبشر قراءتها، وهذه المعلومات ناجمة عن عمليات بحث تستند إلى أنطولوجيا عن كيانات. على سبيل المثال، ستكون نتيجة البحث عن «توني بلير» داخل محرك جوجل عرضًا ملخصًا على يمين الشاشة تظهر فيه عدة صور ومعلومات أساسية، مثل تاريخ الميلاد، وهذه النتائج تُولّد بصورة آلية من التمثيل الرسومي للمعرفة الرسمية الخاصة بتلك الصور والحقائق [307].

النتائج المهيكلة: عادة ما تُعرض عمليات البحث المستندة إلى الأنطولوجيات التي تنتج عنها قائمة من الكيانات في صيغة مهيكلة، على سبيل المثال قائمة تضم أسماء مدن المملكة المتحدة. راجع على سبيل المثال عمليات البحث[1] التي تتم بواسطة نظام KIM [296] أو نظام بروكولي [306].

١- تتوفر مجموعة من الاستعلامات المقدمة كأمثلة وعدد من مؤشرات Mímir التجريبية لغرض إجراء التجارب على الموقع: /http: //demos.gate.ac.uk /mimir

٩-١-٢ لماذا يُستخدم بحث النص الكامل الدلالي؟

تثبت الدراسة [305]، أن عمليات بحث النص الكامل الدلالي تعطي نتائج جيدة في عمليات البحث المهتمة بالدقة، وذلك عندما تتضمن المستندات الكلمات المفتاحية التي تصف حاجة المستخدم. لكن هناك العديد من الحالات التي تكون فيها القدرة على استرجاع المعلومات (recall) ذات أهمية قصوى، وتكون هناك حاجة للحصول على معرفة ضمنية من أجل الرد على أجزاء من الاستعلام. هناك نوع شائع من هذه الاستعلامات، وهو الاستعلام المستند إلى الكيانات، ومن الأمثلة على ذلك «النباتات ذات الأوراق القابلة للأكل» [305]. في هذه الحالة، من المرجح ألاَّ يوجد مستند واحد يحتوي الإجابة، كما تشير المستندات عادة إلى أنواع النباتات المحددة بالاسم (مثل البروكلي)، بدلاً من استخدام مصطلح «نباتات» العام.

العلوم البيئية هي مثال آخر على المجالات التي تكون فيها حاجة قوية للذهاب خطوة أبعد عن عمليات البحث المستندة إلى الكلمات المفتاحية [308، 309]. قامت المكتبة البريطانية بإجراء مسح شمل الباحثين في مجال العلوم البيئية، وأجرت تحليلاً لأنواع احتياجات المعلومات التي واجهوا صعوبة في تلبيتها عبر عمليات البحث بواسطة الكلمات المفتاحية [310]. كان المطلب الرئيس يتعلق بالاستعلامات الخاصة بمنطقة جغرافية معينة، بما فيها البحث المتعلق بالمناطق المجاورة لمنطقة ما (مثال: «مستندات تتعلق بالفيضانات في المناطق التي تبعد ٥٠ ميلاً عن شيفيلد») والمواقع الضمنية (مثال: يجب أن تكون نتيجة الاستعلام «مستندات تتعلق بالفيضانات في المناطق التي تبعد ٥٠ ميلاً عن شيفيلد» مستندًا يتعلق بالفيضانات في مدينة إكستر، على الرغم من أن منطقة جنوب غرب إنجلترا لم يرد ذكرها صراحة).

هناك مثال آخر وهو البحث في براءات الاختراع [295، 311]، حيث تكون القدرة على استرجاع المعلومات بالغة الأهمية، وذلك لأن الإخفاق في العثور على براءات اختراع موجودة مسبقًا وذات صلة قد يؤدي إلى الدخول في مرافعات قضائية وتكبد خسائر مالية. من الأمثلة التي تدل على المعلومات التي يصعب العثور عليها باستخدام الكلمات المفتاحية وحدها عمليات البحث عن إشارات مرجعية إلى أوراق

بحثية مقتبسة في قسم محدد من براءة الاختراع، وكذلك عمليات البحث عن القياسات والكميات (في براءات الاختراع الكيميائية مثلاً). تكون القياسات ذات طبيعة عددية بصفة خاصة، وقد تظهر عليها اختلافات كبيرة -فقد يجري التعبير عن القيمة نفسها باستخدام أنظمة قياس مختلفة كالبوصات أو السنتيمترات أو المضاعفات المختلفة، حتى عند استخدام نظام القياس نفسه كالمليمترات أو السنتيمترات أو الأمتار.

٩-١-٣ استعلامات البحث الدلالية

نظرًا لضرورة أن تتضمن استعلامات البحث الدلالية كلمات دلالية نصية واستعلامات شبيهة باستعلامات لغة «سباركل» (SQARQL) داخل الأنطولوجيا، فعادة ما يُشار إليها بالاستعلامات الهجينة. يستخدم نظام Semplore [312] على سبيل المثال رسوم استعلام رابطة هجينة (conjunctive hybrid query graphs)، تكون شبيهة باستعلامات لغة «سباركل» (SQARQL)، لكنها معززة بمفهوم «افتراضي» يُسمى مفهوم الكلمة المفتاحية W. هناك منهجية أخرى مشابهة جرى اتباعها في نظام Broccoli [306]، ويوجد بها علاقة «occurs – with» (يحدث مع)، وتكون قيمتها الكلمة المفتاحية في النص الحر.

يوجد في نظام Mímir[1] [295] لغة استعلام أكثر ثراءً، كما تدعم هذه اللغة إضافة الشروح اللغوية إلى البحث. على سبيل المثال، تكون نتيجة الاستعلام «شخص يقول» باستخدام نظام Mímir مستندات يوجد داخلها كيانات من نوع «شخص» متبوعة بالكلمة المفتاحية «يقول». كما تدعم الاختلافات النحوية في الكلمات المفتاحية (مثال: «شخص، الجذر:قول»)، وهذا ينطبق أيضًا على قيود المسافة (مثال: «شخص [٠..٥] الجذر: قول»)، حيث تتطابق النتيجة مع كلمات يصل عددها إلى 5 كلمات تفصل بين المكونين، مثل «سيباستيان جيمس من مجموعة ديكسونس قال»). يجري التعبير عن القيود الدلالية الإضافية المبنية على المعرفة المأخوذة من الأنطلوجيا عن طريق إضافة استعلام لغة «سباركل» «SPARQL». على سبيل المثال، يكون هذا الاستعلام للمستندات التي تذكر الأشخاص المولودين في مدينة شيفيلد:

1- http://gate.ac.uk/mimir/

{Person sparql = “SELECT ?inst
WHERE { ?inst :birthPlace <http://dbpedia.org/resource/Sheffield>}

٩-١-٤ تحديد الدرجات واسترجاع البيانات حسب الصلة

في سياق بحث النص الكامل الدلالي، تقترح دراسة [313] إجراء تعديل على tf.idf (تكرار النص. عكس تكرار المستند)، بناءً على تكرار ورود الحالات (instances) من الشروح الدلالية في مجموعة المستندات. كما تجمع بين أوجه الشبه الدلالي مع وجه شبه معياري مبني على الكلمات المفتاحية لإجراء عملية التصنيف، من أجل أخذ الحالات التي لا توجد فيها شروح دلالية على درجة كافية من الصلة في الحسبان.

يدعم إطار عمل بحث النص الكامل الدلالي في نظام Mímir [295] وظائف تصنيف مختلفة، كما يمكن إدراج وظائف جديدة فيه. بالإضافة إلى tf.idf (تكرار النص. عكس تكرار المستند)، يقوم كذلك بتطبيق تصنيف مبني على طول النتائج المطابقة للاستعلام وخوارزمية BM25.

يذهب نظام CE^2 خطوة أبعد من ذلك ويستخدم منهجية مبنية على الرسوم البيانية لحساب تصنيف نتائج البحث الهجينة [314]. يؤخذ هيكل الرسوم البيانية من المعرفة الرسمية الدلالية.

فيما يتعلق بتصنيف الأشخاص الذي ينتج عبر عمليات البحث داخل قواعد المعرفة، تقترح دراسة [315] منهجية ObjectRank وهي منهجية مبنية على تصنيف الصفحة.

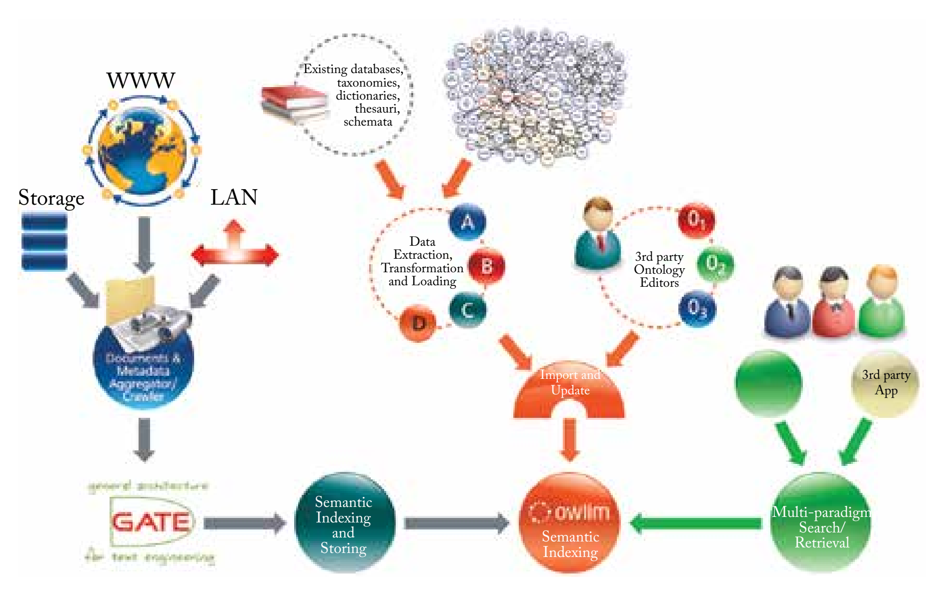
٩-١-٥ منصات بحث النص الكامل الدلالي

سنورد فيما يلي بعضًا من أهم أطر العمل/ النماذج الأولية في البحث الدلالي، مع الإشارة إلى وجود الكثير من أطر العمل أو النماذج الأخرى.

نظام GoNTogle [316] هو نظام بحث يقدم إمكانية البحث بواسطة الكلمات المفتاحية أو الدلالات أو بمنهجية هجينة تجمع بين الكلمات المفتاحية والدلالات داخل مستندات تتضمن شروحًا دلالية. يقوم البحث الدلالي باستبدال الكلمات

المفتاحية بالفئات (classes) الأنطولوجيا. تأتي النتائج حسب ورود فئات الأنطولوجيا الموجودة في الاستعلام في الشروح الخاصة بمستند معين. أخيرًا، يتكون البحث الهجين من عمليات AND أو OR المنطقية (boolean) المعيارية وتطبق على مجموعات النتائج التي يجري توليدها بواسطة بحث بالكلمات المفتاحية وبحث دلالي. النوع الوحيد من الشروحات المدعومة في هذا النظام هو ربط فئة من فئات الأنطولوجيا بجزء من مستند معين. هناك نظام مشابه آخر، وهو نظام Semplore [312] الذي يستخدم رسوم الاستعلام الهجينة الرابطة، مثل نظام SPARQL، لكنها معززة بمفهوم «افتراضي» يُسمى مفهوم الكلمة المفتاحية W. لكن نظام GoNTogle ونظام Semplore لا يدعمان إمكانية البحث في هيكل المستندات، أو إمكانية البحث في أنواع الشروح اللغوية الأخرى.

بدوره يوفر نظام Broccoli [306] واجهة مستخدم لإنشاء الاستعلامات، وذلك بالجمع بين قيود نصية وقيود دلالية (مشفرة كإشارات للكيانات في النص المُدخل، بواسطة معرفات موارد موحدة (URIs)). يُشفر الارتباط بين النص والدلالات بواسطة علاقة occurs-with التي يُشار إليها ضمنيًّا كلما وردت الإشارات إلى الكلمات وكيانات الأنطولوجيات في السياق نفسه. تُستخرج السياقات تلقائيًّا في زمن الفهرسة (indexing time)، وتعتمد في الغالب على التحليل السطحي للمستند واستخراج علاقات التبعية النحوية. توفر علاقة occurs-with القدرة على الوصول إلى هيكل العبارات الكامن في المسند المُدخل. غير أن النظام مصمم ليستخدم فقط هذه العلاقة المحددة، لذا من المرجح أن تكون عملية فهرسة المستندات ذات الهياكل المختلفة (مثال: النبذة المختصرة، الأقسام) صعبة. من ثم لا يوجد دعم لإضافة تعليقات لغوية أكثر ثراءً، مثل أقسام الكلام أو الصرف الإعرابي أو البيانات الوصفية الخاصة بالمستند أو البحث الهيكلي باستثناء البحث الهيكلي المستند إلى التواردات المشتركة داخل السياقات.



الشكل ٩-١: هيكل منصة KIM.

كانت منصة KIM (إدارة المعرفة والمعلومات) [296، 317] من بين أوائل الأنظمة التي طبقت البحث الدلالي، سواءً أكان داخل قواعد RDF المعرفية بواسطة لغة سباركل «SPARQL» أم داخل محتوى المستندات التي تتضمن الشروحات الدلالية، بما في ذلك الاستعلامات الهجينة التي تخلط بين الكلمات المفتاحية والقيود الدلالية. يوجد في منصة KIM عدد من واجهات المستخدم الخاصة بالبحث الدلالي والتصفح، ويمكن تكييفها بسهولة لتتناسب مع تطبيقات محددة. هذا النظام متوفر للاستخدامات البحثية عبر الرابط http: / /www.ontotext.com /kim /getting-started /download.

منصة KIM هي منصة قابلة للتمديد لإدارة المعرفة، حيث توفر أدوات لإضافة الشروحات الدلالية والفهرسة وإجراء عمليات البحث استنادًا إلى الدلالات (يُشار إليها باسم البحث متعدد الجوانب في منصة KIM). يظهر الشكل رقم 1-9 هيكل منصة KIM التي تتضمن كذلك جامع بيانات الويب (web crawler) لجمع المحتوى، ووحدة استخراج المعرفة وتحويلها وتحميلها (ETL) تكون بمنزلة رابط يربط بموسوعات المفردات والقواميس وموارد LOD، إضافة إلى مجموعة من واجهات المستخدم مبنية على شبكة الإنترنت لإجراء عمليات البحث باستخدام الكيانات أو

الدلالات (راجع القسم ٩-١-٦ لمعرفة تفاصيل البحث متعدد الجوانب باستخدام منصة KIM).

تعتمد إضافة الشروحات الدلالية في منصة KIM على أدوات معالجة اللغات الطبيعية في منصة GATE. يتمثل جوهر عملية إضافة الشروحات الدلالية في منصة KIM على التعرف على كيانات الأسماء ذات الصلة بأنطولوجيا KIM. تحمل جميع حالات الكيانات مُعرّفات فريدة تسمح بربط الشروحات بنوع الكيان والشخص المحدد في قاعدة الحالات. تُخصص مُعرّفات جديدة للكيانات الجديدة (غير المعروفة سابقًا)، وبعدها تُضاف أوصاف محدودة إلى المستودع الدلالي. تُحفظ الشروحات بصورة منفصلة عن المحتوى، وتُقدم واجهة برمجة تطبيقات (API) لإدارتها.

يمكن لمنصة KIM كذلك استخدام أنطولوجيات البيانات المترابطة لغرض إضافة التعليقات الدلالية وإجراء الأبحاث الدلالية. في الوقت الحالي، جرى اختبارها مع قواعد من بينها DBPedia وGeonames وWordnet وMusicbrainz وFreebase وUMBEL وLingvoj وكتاب حقائق العالم الذي تصدره وكالة المخابرات الأمريكية. تُعالج مجموعات البيانات هذه بصورة مسبقة وتُشغّل لإنشاء مجموعة بيانات متكاملة تضم نحو ١,٢ مليار عبارة صريحة. تُجرى أيضًا عملية التسلسل الأمامي (forward-chaining) لبلورة ٠,٨ مليار عبارة ضمنية إضافية.

GATE Mímir[1] [295] هو إطار عمل متكامل لإجراء عمليات البحث الدلالي، ويتيح الفهرسة والبحث داخل النص الكامل وهياكل المستندات والبيانات الوصفية الخاصة بالمستندات والشروحات اللغوية وأي قواعد معرفة مترابطة خارجية. كما يدعم الاستعلامات الهجينة التي تمزج بصورة عشوائية بين النص الكامل والقيود الهيكلية واللغوية والدلالية. هناك ميزة أساسية تميزه عن الأعمال السابقة، وهي معاملات الاحتواء (containment operators) التي تسمح بإنشاء قيود النص الكامل والقيود الهيكلية والدلالية بمرونة، وجعل هذه القيود متداخلة.

1 http://gate.ac.uk/projects/envilod

يبين الشكل ٩-٢ واجهة المستخدم الخاصة بالاستعلامات الدلالية في نظام Mímir. يتمثل الهدف في العثور على مستندات يرد فيها ذكر مواقع في المملكة المتحدة تكون فيها الكثافة السكانية أكثر من ٥٠٠ شخص في الكيلومتر الواحد. تأتي المعرفة بالكثافة السكانية من قاعدة DBpedia. تكون المستندات التي يجري البحث فيها بيانات وصفية للتقارير الحكومية الخاصة بالتغير المناخي والفيضانات أنشأتها المكتبة البريطانية كجزء من مشروع EnviLOD [١].

يتمثل المفهوم العام الذي يقوم عليه نظام Mímir في أن مجموعة المستندات تُعالج بواسطة خوازرميات معالجة اللغات الطبيعية، وعادة ما تتضمن عملية المعالجة إضافة الشروحات الدلالية باستخدام البيانات المترابطة المفتوحة التي يتم الوصول إليها عبر إحدى قواعد كيانات البيانات الثلاثية (triplestore)، مثل OWLIM [318] أو Sesame. بعدها تجري فهرسة المستندات التي أضيفت إليها الشروحات في نظام Mímir، إلى جانب محتوى النص الكامل الخاص بها والبيانات الوصفية الخاصة بالمستند وعلامات هيكل المستند (يمكن اكتشاف علامات هيكل المستند بشكل آلي بواسطة أدوات معالجة اللغات الطبيعية). أثناء إجراء البحث، تُستخدم قاعدة كيانات البيانات الثلاثية كمصدر للمعرفة الضمنية، وذلك للمساعدة في الإجابة عن الأبحاث الهجينة التي تجمع بين النص الكامل والقيود الهيكلية والدلالية. تُنشأ القيود الدلالية باستخدام استعلام لغة «سباركل» (SPARQL) وتُطبق على قاعدة كيانات البيانات الثلاثية.

يستخدم نظام Mímir فهارس مقلوبة لفهرسة محتوى المستند (بما في ذلك المعلومات اللغوية الإضافية كأقسام الكلام أو الجذور الإعرابية)، وللربط بين حالات الشروحات مع الموقع الذي توجد فيه داخل النص المُدخل. الفهرس المقلوب المستخدم في نظام Mímir مبني على محرك MG4J [319]. إضافة إلى نص الوثيقة، النوع الرئيس الآخر من البيانات هو الشروحات الهيكلية والشروحات المولدة بواسطة مهام معالجة اللغات الطبيعية. في نظام Mímir، يوجد تمثيل لكلا النوعين داخل هيكل البيانات نفسه،

١- متاحة أون لاين عبر http://exopatent.ontotext.com

ويتألف من موقع بدء وموقع نهاية، ونوع الشرح (مثال: موقع) ومجموعة اختيارية من الخصائص (تسمى السمات في إطار عمل GATE).

نظام Mímir قابل للتوسيع بشكل كبير، ففي أحد التطبيقات جرت فهرسة 150 مليون صفحة ويب بنجاح، باستخدام مئتي عنصر كبير لأمازون Amazon) EC2 EC2 Large Instances) والتي ظلت تعمل لمدة أسبوع من أجل توليد فهرسة موحدة [293]. نظرًا لكون نظام Mímir يعمل بواسطة منصة GateCloud.net لمعالجة النصوص [320]، فإن عملية بناء الفهارس الدلالية في سحابة أمازون هي عملية سهلة.

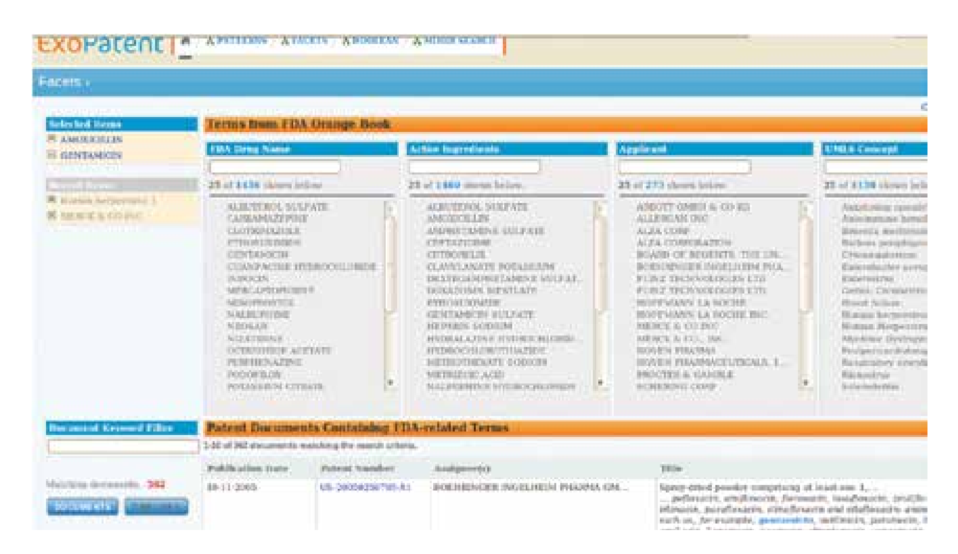
الشكل ٩-٢: واجهة المستخدم الخاصة بالبحث الدلالي في نظام Mímir يظهر فيها استعلام رسمي والوثائق المسترجعة ومقتطفات نصية قصيرة تظهر المواقع المطابقة للاستعلام بالخط العريض.

٩-١-٦ البحث متعدد الجوانب المستند إلى الأنطولجيا

كما سبق أن ناقشنا، توجد في نظام KIM مجموعة شاملة من واجهات المستخدم المستندة إلى متصفحات الويب لإجراء عمليات البحث الدلالية. يشمل ذلك البحث المتعدد الجوانب المعتمد على الأنطولوجيا، حيث يستطيع المستخدم اختيار حالة واحدة أو أكثر (مجسدة في شكل صور بواسطة ملصقات RDF الخاصة بها، لكن العثور عليها

يكون بواسطة مُعرّفات الموارد الموحدة (URIs) الخاصة بها) والحصول على مستندات ترد فيها بشكل مشترك. كما يدعم النظام العرض بالخط الزمني أو بشكل متمحور حول الكيانات.

يبين الشكل ٩-٣ حالة يبحث فيها المستخدم عن براءات اختراع تذكر دوائي أموكسيسلين وجينتمايسين. هذا المثال مأخوذ من النسخة الإلكترونية التجريبية لنظام KIM[1] في ExoPatent، التي تستخدم كتاب إدارة الغذاء والدواء الأصفر (يضم ٢٣,٠٠٠ دواء حاصل على براءة اختراع) ونظام اللغة الطبية الموحد (UMLS -قاعدة بيانات مؤلفة من ٣٧٠,٠٠٠ مصطلح طبي) لإضافة المعلومات الدلالية كشروحات إلى المستندات. تعمل النسخة التجريبية على مجموعة صغيرة من براءات الاختراع يصل عددها إلى ٤٠,٠٠٠. يدعم نظام ExoPatent البحث الدلالي عن الأمراض وأسماء الأدوية وأعضاء الجسم والإشارات إلى الأدبيات وبراءات الاختراع الأخرى والقيم العددية والنطاقات.



الشكل ٩-٣: واجهة المستخدم الخاصة بالبحث المتعدد الجوانب المستند إلى الكيانات في نظام KIM.

1- http://ideya.eu.com/reports.html

في واجهة المستخدم الخاصة بالبحث المتعدد الجوانب، يجري تحديث عدد المستندات المطابقة بصورة ديناميكية بالتزامن مع اختيار الكيانات الجديدة كقيود (انظر العمود على يسار الشكل). كما يمكن تحديد قيود الكلمات المفتاحية الاختيارية داخل حقل الفلتر الموجود على اليسار. في أسفل الشكل، يمكن رؤية عنوانات المستندات المسترجعة وبعض المحتويات ذات الصلة الموجودة فيها. عنوانات المستندات قابلة للضغط من أجل الاطلاع على المحتوى الكامل في المستند والتعليقات الدلالية داخله. يجري أيضًا تحديث الكيانات/ المصطلحات المدرجة في عمود الكيانات (اسم الدواء والمكونات وصاحب الطلب ومفهوم نظام اللغة الطبية الموحد) لإظهار الكيانات المتواردة بشكل مشترك مع قيود الكيانات المختارة مسبقًا فقط.

يوجد في نظام Broccoli المذكور سابقًا واجهة مستخدم تفاعلية مشابهة لإنشاء الاستعلامات، حيث يجري تحديثها بصورة آلية بالتزامن مع كتابة المستخدم المفاهيم أو الكلمات المفتاحية التي يرغب في البحث عنها. تكون المستندات التي يجري البحث فيها مقالات ويكيبيديا مفهرسة بواسطة الفئات (classes) والحالات (instances) المأخوذة من أنطولوجيا YAGO. يبين الشكل ٩-٤ استعلامًا يقدم كمثال للمستندات التي تذكر المدن البريطانية التي تتضمن أيضًا الكلمة المفتاحية «فيضان». يُعرض الاستعلام الدلالي كرسم بياني في القسم العلوي، وهو ما يجعل العلاقات القائمة بين المفاهيم التي يجري البحث عنها صريحة. تمتلك الكلمات المفتاحية علاقة خاصة هي علاقة «occurs-with»، في حين تأتي جميع العلاقات الدلالية الأخرى من أنطولوجيا YAGO. مع بدء المستخدم كتابة مصطلح استعلام (مثال: مدينة)، يجري تحديث قوائم الفئات (classes) والحالات (instances) والعلاقات (relations) المطابقة الموجودة على اليسار بصورة ديناميكية. بعد اختيار مصطلح استعلام، لا يجري عرض سوى العلاقات المنطبقة على هذه الفئة في قائمة العلاقات المحتملة. بسبب الاستعلامات المتمركزة حول الكيانات، تجري هيكلة قائمة النتائج كقائمة كيانات، حيث تقدم معلومات ذات صلة من أنطولوجيا YAGO لكل كيان يجري عرضه، وكذلك وثائق من موسوعة ويكيبيديا عن هذا الكيان تحتوي كذلك على الكلمة/ الكلمات المفتاحية المعطاة.



الشكل ٩-٤: واجهة Broccoli التفاعلية لإنشاء الاستعلامات.

٩-١-٧ واجهات البحث الدلالي المستندة إلى النماذج

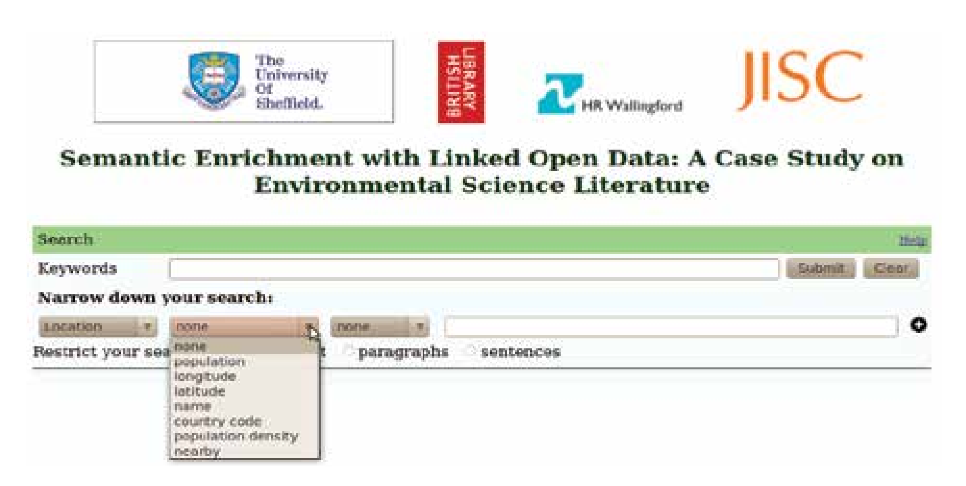
إحدى التحديات التي تواجهها واجهات البحث الدلالي، ولا سيما في الحالات ذات الموضوعات المحددة، هو توضيح ما يمكن البحث عنه للمستخدمين. تجعل واجهات البحث المستندة إلى النماذج هذا الأمر صريحًا، وذلك بصورة تشبه واجهات المستخدم متعددة الجوانب التي ورد نقاشها أعلاه.

يظهر مثال للواجهات المستندة إلى النماذج في الشكل ٩-٥ من واجهة EnviLOD UI [308] التي جرى تطويرها كواجهة أمامية سهلة الاستخدام لإجراء عمليات البحث الدلالي لفهرس Mímir يضم مستندات ومصطلحات وكيانات LOD في مجال العلوم البيئية.

هناك حقل للكلمات المفتاحية، تكمّله قيود اختيارية لإجراء البحث الدلالي، عبر مجموعة من القوائم المنسدلة المعتمد بعضها على بعض. في القائمة الأولى، يستطيع المستخدمون البحث عن أنواع كيانات معينة (المواقع، المؤسسات، الأشخاص، الأنهار، التواريخ)، ويمكنهم كذلك تحديد القيود في الخصائص على مستوى المستند. يمكن كذلك إضافة أكثر من قيد دلالي واحد، وذلك بواسطة زر الإضافة، الذي يقوم بإضافة خانة جديدة تحت خانة القيود الحالية.

على سبيل المثال، في حال اختيار «موقع» كقيد دلالي، يمكن بعدها تحديد قيود إضافية عن طريق اختيار قيد خاصية مناسب، كما هو مبين في الشكل. يسمح القيد «سكان» للمستخدمين فرض قيود على عدد السكان في المواقع التي يجري البحث عنها. يمكن كذلك فرض قيود عددية مشابهة على قيم الارتفاع والطول والكثافة السكانية.

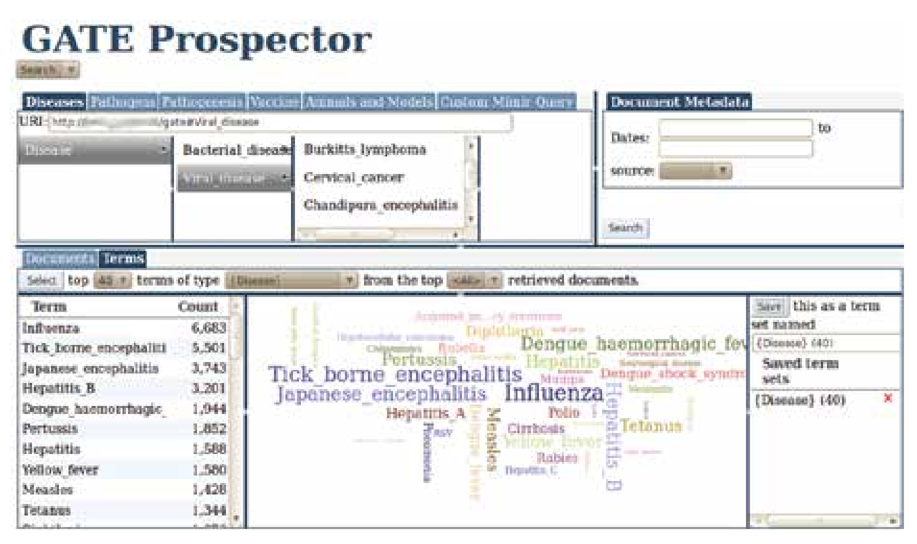
يمكن كذلك فرض القيود من ناحية اسم الموقع أو البلد الذي ينتمي إليه. فيما يتعلق بالخصائص ذات القيم التسلسلية، يجري اختيار كلمة «هو» من القائمة الثالثة بدلاً من «لا شيء»، وبعدها يجب أن تكون القيمة مثلما جرى تحديده تمامًا (مثال: أكسفورد)، في حين تؤدي كلمة «contains» [يحتوي على] إلى التطابق مع سلسلة فرعية من الحروف، (مثال: يتطابق الاستعلام مع كلمة Oxfordshire كاسم موقع يحتوي على كلمة Oxford). بهذه الطريقة، لا يُعرض على المستخدم الذي يبحث عن مستندات تذكر المواقع التي تحتوي على اسم يضم كلمة «Oxford» المستندات التي تذكر كلمة «Oxford» بصورة صريحة فحسب، بل أيضًا المستندات التي تذكر كلمة Oxfordshire والمواقع الأخرى الموجودة في Oxfordshire (على سبيل المثال: وايثام وودز (Wytham Woods)، بانبري (Banbury)). في المثال الأخير، تُستخدم المعرفة المأخوذة من قاعدتي DBpedia وGeoNames لتحديد المواقع الأخرى الموجودة في Oxfordshire، بالإضافة إلى Oxford.



الشكل ٩-٥: واجهة المستخدم الخاصة بالبحث الدلالي في نظام EnviLOD

إحدى المشكلات الموجودة في واجهة المستخدم التي تكون على نمط EnviLOD كونها تُخفي عن المستخدم المعلومات المتعلقة بحالات هذه الفئات التي ترد في مجموعة الوثائق المفهرسة (مثال: مقاطعات المملكة المتحدة المذكورة). من المنهجيات المستخدمة لتوفير هذا النوع من النظرات العامة على المستندات بالاعتماد على الكيانات، إعداد قائمة لجميع الحالات، لكل فئة من الفئات، كما هو الحال في الواجهتين الموجودتين في نظامي KIM و Broccoli.

هناك خيار بديل، وهو استخدام سحابات البطاقات التصنيفية (tag clouds) وغيرها من أساليب تجسيد التواردات المشتركة للكيانات في صيغة مرئية. جرى في الآونة الأخيرة إضافة واجهة مستخدم من هذا النوع إلى نظام Mímir، وتسمى GATE Prospector (راجع الشكل ٩-٦). يُظهر النصف العلوي من واجهة المستخدم فئات وحالات الأنطولوجيا (نظام اللغة الطبية الموحد (UMLS) في هذه الحالة) وبعدها يقوم المستخدم باختيار الفئات والحالات التي يرغب فيها المستخدم. يمكن أيضًا فرض قيود إضافية على البحث عبر فلاتر البيانات الوصفية الخاصة بالوثيقة. يُظهر النصف العلوي من الصورة الحالات المطابقة (أي المصطلحات في حالة نظام اللغة الطبية الموحد (UMLS))، بالإضافة إلى عدد المرات التي ترد فيها في مجموعة الوثيقة. تُعرض أيضًا سحابة مصطلحات مبنية على أساس التكرار. يمكن حفظ مجموعة المصطلحات/ الحالات لاستخدامها لاحقًا، على سبيل المثال لتوليد تجسيدات مرئية للتوارد المشترك بين الكيانات/ المصطلحات.



الشكل ٩-٦: واجهة المستخدم الخاصة بالبحث الدلالي في باحث نظام GATE.

يبين الشكل ٩-٧ مثالاً للتصور البياني للتوارد المشترك، حيث تُرسم نماذج الأمراض الأكثر ذكرًا مقابل نماذج مسببات الأمراض الأكثر ذكرًا. تشمل الأمثلة في النطاقات الأخرى رسم مصطلحات المشاعر التي ترد بصورة هي الأكثر تكرارًا مع الأحزاب السياسية أو السياسيين، في ضوء مجموعة ضخمة من التغريدات المتعلقة بانتخابات معينة.

٩-١-٨ البحث الدلالي في محتوى وسائل التواصل الاجتماعي

يختلف البحث في محتوى وسائل التواصل الاجتماعي بصورة كبيرة عن البحث في شبكة الإنترنت [321] بعدد من الطرق المهمة. أولاً، يبحث المستخدمون داخل تدفقات الرسائل، مثل رسائل تويتر، عن معلومات ذات صلة من الناحية الزمنية، وهم مهتمون بالأشخاص أكثر من أي شيء آخر. ثانيًا، تُستخدم عمليات البحث لمراقبة محتوى تويتر بمرور الوقت، ويمكن حفظها كجزء من ملفات المستخدمين. ثالثًا، تتميز عمليات البحث في تويتر بكونها أقصر بكثير، وبأنها تؤدي إلى قدر أكبر من الإثارة الاجتماعية، في حين تبحث عمليات البحث في الإنترنت عن الحقائق. بالإضافة إلى قِصر الرسالة وطبيعتها المشوشة والمعلومات الإضافية المخفية في روابط الـURL وعلامات الهاشتاغ،

فإن هذه الاختلافات تجعل أساليب البحث التقليدية بواسطة الكلمات المفتاحية دون المستوى الأمثل عندما تُستخدم للبحث في محتوى وسائل التواصل الاجتماعي.

تُظهر مقارنة بين أدوات مراقبة وسائل التواصل الاجتماعي أجريت في أكتوبر ٢٠١٤ من قبل شركة Ideya المحدودة(١) أن هناك ما لا يقل عن ٢٤٥ أداة لمراقبة وسائل التواصل الاجتماعي، منها ١٩٧ أداة مدفوعة، مع كون بقية الأدوات مجانية أو تعمل بنظام يدعى الفريميوم (freemium). غالبية الأدوات المجانية، على الأقل، لا تسمح بإجراء التحليل المتعمق والقابل للتخصيص المطلوب من الناحية المثالية. ركزت الأبحاث المنشورة بشكل رئيس على التمرينات التي تقوم بإجراء عمليات حسابية بناءً على تمييز الموضوع والهوية بواسطة علامات الهاشتاغ والكلمات المفتاحية البسيطة أو البيانات الوصفية الخاصة بتويتر المتاحة بسهولة، كاسم المؤلف واللغة وعدد مرات إعادة التغريد وما شابه [326–322]. في حين تتضمن بعض من هذه الأساليب أدوات أكثر تعقيدًا للقيام بمهام المعالجة اللغوية، لكنها عادة ما تتكون من أدوات بسيطة جاهزة لتحليل المشاعر، مثل أداة SentiStrength [214] وأداة SentiWordNet [327] و/ أو أدوات التعرف على الكيانات والموضوعات العمومية الأساسية مثل أداة DBpedia Spotlight [115]، أو أدوات معالجة اللغات الطبيعية الأساسية مفتوحة المصدر مثل أداة ANNIE [328]، ولا يجري تكييفها مع النطاق والمهمة. لذلك سيركز هذا القسم على الأعمال التي جرت في الآونة الأخيرة وتناولت البحث الدلالي وتهدف إلى معالجة هذه التحديات.

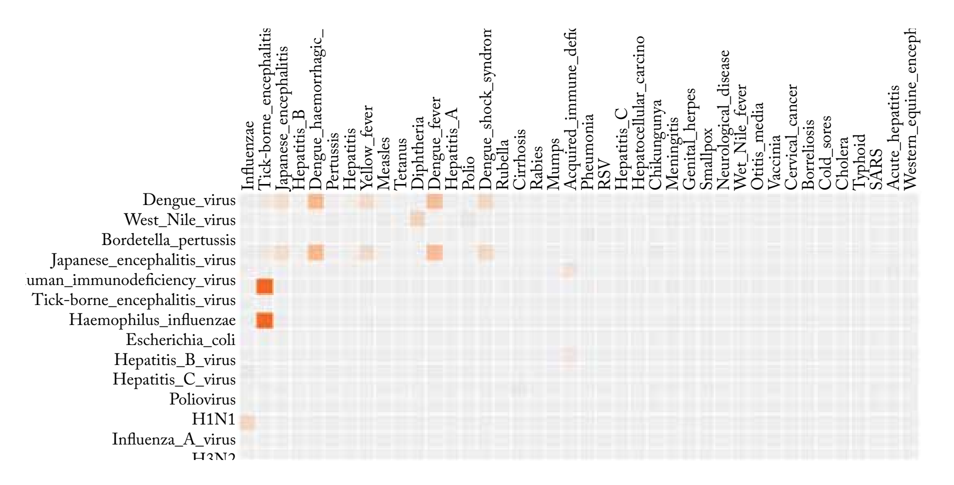
أعطى مؤتمر استرجاع المعلومات ٢٠١١ لمراقبة المدونات المصغرة (TREC 2011 Microblog track)(٢) زخمًا جديدًا للأبحاث عن طريق توفير مجموعة من موضوعات الاستعلامات، ونقطة زمنية، ومكنزًا يضم ١٦ مليون تغريدة، منها مجموعة فرعية أضيفت إليها شروحات بشكل يدوي لتحديد الصلة كمعيار ذهبي. بالإضافة إلى الخصائص المستخدمة على نطاق واسع المستندة إلى الكلمات المفتاحية وخصائص

1- http://sites.google.com/site/trecmicroblogtrack/

٢- لمزيد من المعلومات، راجع https://gate.ac.uk/gcp/

التركيب النحوي للتغريدات (مثال: ما إذا كانت التغريدات تحتوي على علامات الهاشتاغ)، أجرى (تأو وآخرون) [329] تجربة على الخصائص النحوية المستندة إلى الكيانات المولّدة بواسطة أداة DBpedia Spotlight، وهو ما يُعطي نتائج أفضل بكثير.

يقوم نظام Twarql [330] بتوليد ثلاثيات RDF من التغريدات، اعتمادًا على البيانات الوصفية المأخوذة من التغريدات نفسها، بالإضافة إلى الإشارات إلى الكيانات وعلامات الهاشتاغ وروابط URL [221]. تُشفّر هذه المعلومات باستخدام مصطلحات Open Data (FOAF, SIOC) المعيارية (راجع القسم ٨-٢) ويمكن بحثها عن طريق استعلامات لغة سباركل (SPARQL). يمكن أيضًا الاشتراك في سلسلة من التغريدات المطابقة لاستعلام دلالي معقد، مثل المنافسين المذكورين مع منتجي (جهاز آيباد من شركة أبل في حالة الاستخدام الخاصة بها). حتى الانتهاء من تأليف هذا الكتاب، لم يُقيّم نظام Twarql بشكل رسمي، وهو ما يعني أن فعاليته ودقته لم تؤكد بعد.



الشكل ٩-٧: باحث نظام GATE: شاشة عرض التوارد المشترك بين الحالات/ المصطلحات.

يقترح (أبيل وآخرون) إطار عمل تكيّفي متعدد الجوانب لإجراء عمليات البحث لتدفقات وسائل التواصل الاجتماعي [331]. يستخدم هذا النظام شروحات الكيانات الدلالية الموجودة في OpenCalais، بالإضافة إلى نموذج مستخدم (راجع القسم ٩-٢-١)، من أجل إنشاء الجوانب (facets) وتصنيفها دلاليًّا. تُستخدم عمليات البحث بواسطة الكلمات المفتاحية والجوانب (facets) المستندة إلى علامات الهاشتاغ

كخطين أساسيين. تُحقق أفضل النتائج عندما تكون الجوانب (facets) ذات طابع شخصي وعندما تُصنف وفقًا للكيانات التي تكون ذات أهمية بالنسبة للمستخدم المعني (كما هو مُشفّر في نموذج المستخدم المستند إلى الكيانات). يتعين أن تكون عملية تصنيف الجوانب (facet) حساسة للسياق الزمني (أي الفرق بين وقت الاستعلام والختم الزمني للنشر).

هناك أيضًا إطار عمل مبني على أساس منصة GATE لتحليل كميات ضخمة من محتوى وسائل التواصل الاجتماعي وبحثها. يتكون إطار العمل هذا والذي يعمل في الوقت الحقيقي (real time) من مكونات إضافة الشروحات الدلالية التي ورد نقاشها في الفصول السابقة، بالإضافة إلى إطار عمل Mímir للبحث الدلالي، ومكون يقوم بتجميع النتائج بشكل ديناميكي. يدعم الإطار البحث الاستكشافي وبناء المعنى عبر واجهات عرض المعلومات في صيغة صور (information visualization interfaces)، مثل مقاييس التوارد المشترك (co-occurrence matrices) وسحابات المصطلحات (term clouds) وخرائط الأشجار (treemaps) وخرائط كوروبليث (choropleths). كما توجد واجهة تفاعلية للبحث الدلالي مبنية على الباحث (Prospector)، حيث يستطيع المستخدمون حفظ نتائج استعلامات البحث الدلالي وتنقيحها وتحليلها بمرور الوقت. جرت برهنة وجود استخدامات عملية لإطار العمل في الزمن الحقيقي وعلى نطاق واسع عبر إجراء تحليل لتغريدات سياسيين بريطانيين وردود الجمهور العام عليهم خلال الفترة التي سبقت الانتخابات العامة التي جرت في المملكة المتحدة في عام 2015، وعبر تحليل أكثر من ٦٤ مليون تغريدة ذات صلة بالاستفتاء الذي جرى في المملكة المتحدة في عام ٢٠١٦ حول عضوية البلاد في الاتحاد الأوروبي (البريكسيت).

بإمكان إطار العمل المستند إلى منصة GATE تنفيذ جميع الخطوات في عملية التحليل، وهي جمع البيانات وإضافة الشروحات الدلالية والفهرسة والبحث وعرض النتائج في صيغة صور مرئية. خلال عملية جمع البيانات، يمكن متابعة حسابات المستخدمين وعلامات الهاشتاغ عبر واجهة برمجة تطبيقات «الحالات/ الفلتر» في تويتر.

يؤدي ذلك إلى توليد ملف مكتوب بلغة JSON يُحفظ لإجراء عملية معالجة في وقت لاحق. يمكن كذلك تحليل تدفقات التغريدات (اختياريًّا) مع وصولها تباعًا، وذلك بشكل آني تقريبًا، وتجري فهرسة النتائج لغرض تجميعها والبحث فيها وعرضها في صيغة مصورة. تُستخدم مكتبة العميل «hosebird» الخاصة بتويتر لإتمام الاتصال بواجهة برمجة التطبيقات، مع إمكانية إعادة الاتصال وإجراء عملية التراجع وإعادة المحاولة (backoff-and-retry) بصورة آلية.

في حالة المعالجة غير المباشرة (non-live processing)، تجري معالجة ملف JSON باستخدام أداة (GATE Cloud Parallelizer)، وهي عبارة عن أداة موازاة سحابة منصة GATE (GCP) لتشغيل ملفات JSON كمستندات GATE (مستند واحد لكل تغريدة) وإضافة الشروحات إليها ومن ثمّ فهرستها لتمكين إجراء البحث والعرض في صيغة الصور في إطار عمل Mímir التابع لمنصة GATE [295]. أداة GCP هي أداة مصممة لدعم تنفيذ منظومات مهام GATE باستخدام مجموعات ضخمة تضم ملايين المستندات، وباستخدام هيكل هندسي متعدد الخيوط[1]. تحدد مهام أو مجموعات أداة GCP باستخدام لغة XML، حيث يُوصَف موقع وصيغة الملفات المُدخلة، وتطبيق GATE الذي ينبغي تشغيله، وأنواع المُخرجات المطلوبة. تُوفر عدد من أدوات مناولة صيغ البيانات المُخرجات (مثل XML وJSON)، لكن جميع المكونات المختلفة هي قابلة للتوصيل (pluggable)، لذا يمكن استخدام طرق تنفيذ خاصة إن كانت المهمة تتطلب ذلك. تحفظ أداة GCP تَقدُّم كل مجموعة في صيغة XML قابلة للقراءة من قبل البشر والآلات. صممت الأداة بصورة تتيح إمكانية إعادة تشغيل مجموعة توجد قيد التشغيل بالإعدادات نفسها إن طرأ عطل عليها لأي سبب من الأسباب، حيث تستأنف أداة GCP العمل بصورة آلية من المكان الذي توقفت عنده.

في الحالات التي يكون من المطلوب إجراء تحليل آني للتدفقات المباشرة، يُستخدم برنامج تدفقات تويتر لإضافة التغريدات الواردة إلى طابور رسائل. بعدها تقوم عملية

---

١ – تشير «SNP Other» الحالة الغريبة التي لم يكن فيها الحزب الوطني الاسكتلندي يشغل المقعد البرلماني أو يتنافس عليه مرشح من الحزب، لكن مع ذلك كان للحزب أهمية تستحق أن نقوم بمتابعته. تشير «Other MP» إلى نواب برلمانيين آخرين ينتمون إلى الأحزاب السياسية الصغيرة الأخرى.

منفصلة لإضافة الشروحات الدلالية (أو عدة عمليات) بقراءة الرسائل من الطابور وتحليلها ودفع الشروحات والنصوص الناتجة إلى Mímir. إن تجاوز معدل التغريدات الواردة الطاقة الاستيعابية لجهة المعالجة، تُطلق حالات إضافية من مستهلك الرسائل عبر آلات متعددة لتوسيع نطاق الطاقة الاستيعابية.

يتكون نظام المعالجة المباشرة من عدة مكونات متمايزة:

مكون الجمع يتلقى التغريدات من موقع تويتر عبر واجهة برمجة التطبيقات (API) streamingومن ثم يقوم بتمريرها نحو طابور رسائل موثوق. كما يقوم بحفظ ملف JSON غير المعالج الذي يحتوي التغريدات في ملفات احتياطية لغرض إجراء المعالجة في وقت لاحق إن دعت الحاجة لذلك.

يستهلك مكون المعالجة التغريدات الموجودة في طابور الرسائل ويقوم بمعالجتها مع منظومة التحليل في منصة GATE ويرسل المستندات التي أضيفت إليها التعليقات إلى نظام Mímir لغرض الفهرسة.

يتلقى نظام Mímir التغريدات التي أضيفت إليها التعليقات ويقوم بفهرسة نصها وبيانات الشروح، ويجعلها متاحة للبحث بعد تأخير قصير (قابل للتهيئة).

بمجرد إضافة التعليقات الدلالية إلى التغريدات وتخزينها في نظام Mímir لغرض إجراء البحث، باستطاعتنا استخدام الباحث (Prospector) لاستعلام نتائج البحث الدلالي وعرضها في صيغة مصورة. في هذا المثال، تُحول مجموعتان من التعليقات الدلالية (الموضوعات السياسية مقابل الأحزاب السياسية البريطانية في هذه الحالة) إلى مصفوفة ثنائية الأبعاد، في حين تعبر شدة لون كل خلية مدى قوة التوارد المشترك. يمكن إعادة تنظيم المصفوفة بالضغط على أي خانة أو عمود، وهو ما يؤدي إلى تصنيف المحور حسب قوة الارتباط مع العنصر الذي جرى الضغط عليه. هذا المثال يعرض الموضوعات العشرة التي جرى التحدث عنها بالصورة الأكثر تكرارًا خلال المرحلة التي سبقت الانتخابات البريطانية التي جرت في عام ٢٠١٥ من قبل أكثر عشر مجموعات قامت بنشر تغريدات، حيث تمثل المجموعة الواحدة حزبًا أو فئة سياسية

(عضو أو مرشح برلماني)[1].

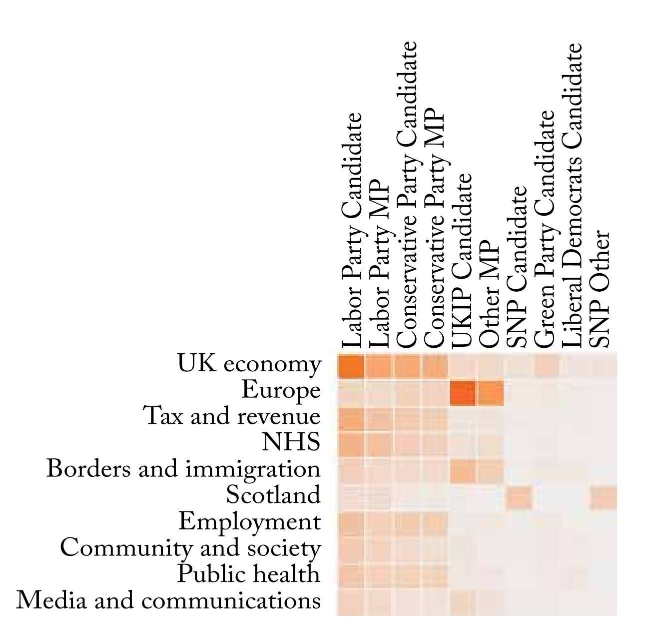
استعلام Mímir الكامن الذي يحدد الموضوعات التي ذُكرت من قبل كل حزب من الأحزاب المشاركة في تغريدات الانتخابات هي كالتالي:

{DocumentAuthor author_party =

"Green Party"}| OVER

{Topic theme = "uk_economy"{

تُضاف المعلومات المتعلقة بالحزب الذي ينتمي إليه ناشر التغريدات والمصطلحات الواردة في كل تغريدة تلقائيًا من قاعدة بيانات DBpedia أثناء مرحلة إضافة التعليقات الدلالية.



الرسم ٩-٨: مصفوفة الباحث (Prospector) للتوارد المشترك بين الموضوع وحزب المرشح.

١- نحو ٣٦٪ من المستخدمين قاموا فعليًا بتعبئة معلومات موقعهم داخل حساباتهم مع توفير المكان الصحيح عبر تحديد أقرب مدينة لهم [٣٤٢].

٩-٢ نمذجة المستخدم المستندة إلى الدلالات

هناك مجال آخر من مجالات تطبيقات بحوث الويب الدلالي التي تستخدم تقنيات معالجة اللغات الطبيعية بشكل كبير، وهو مجال النمذجة الدلالية للمستخدمين والمجتمعات، ومن الأمثلة على الدراسات التي تتناول ذلك دراستا [134، 332]. نشير هنا إلى أن مراجعة نمذجة المستخدم بشكل مفصل لأغراض الويب الدلالي تتجاوز نطاق هذا الفصل، لكن ننصح بقراءة دراسة [333].

لو تحدثنا بتفصيل أكبر، نمذجة المستخدم (UM) هي مورد معرفة يضم معلومات دلالية صريحة عن جوانب مختلفة تتعلق بالمستخدم، وهذه المعلومات متوفرة بصورة مسبقة (مأخوذة من البيانات الوصفية في حسابات الفيسبوك مثلاً) أو تُستنبط تلقائيًّا من سلوك المستخدم أو من المحتوى المقدم من طرف المستخدمين أو من شبكات التواصل الاجتماعي أو غيرها من المصادر. في العادة تُستخدم أساليب معالجة اللغات الطبيعية كعملية تمييز كيانات الأسماء وربطها لإتمام المهمة الأخيرة.

يتمثل الأساس المنطقي الذي تعتمد عليه عملية اشتقاق نموذج المستخدم بناءً على الأنطولوجيات بصورة آلية من البيانات الاجتماعية في أنها تشكل أساس إدارة المعلومات الشخصية (PIM) اعتمادًا على الدلالات وغيرها من التطبيقات المشابهة. على وجه الخصوص، يعود أصل الأعمال المتعلقة بإدارة المعلومات الشخصية إلى الأبحاث التي أجريت على سطح المكتب الدلالي الاجتماعي (social semantic desktop) [334]، حيث يجري تحليل المعلومات المأخوذة من جهاز الحاسوب المكتبي الخاص بالمستخدم (البريد الإلكتروني أو المستندات مثلاً) بواسطة أساليب معالجة اللغات الطبيعية من أجل اشتقاق نماذج المستخدم.

٩-٢-١ بناء نماذج مستخدم دلالية اجتماعية مأخوذة من الشروح الدلالية

من بين الأنواع المختلفة لوسائل التواصل الاجتماعي، حظيت فهارس المستخدمين (folksonomies) على الأرجح بأكبر قدر من اهتمام الباحثين الذين يقومون بدراسة كيفية اشتقاق نماذج دلالية تعبر عن تفاعلات المستخدمين واهتماماتهم من المحتوى

الذي يقوم المستخدمون بإنتاجه. ركزت العديد من المنهجيات على استكشاف الرسوم البيانية الاجتماعية ورسوم التفاعلات، وذلك باستخدام أساليب مأخوذة من تحليل الشبكات الاجتماعية (مثال: [335]). لكننا في هذا القسم مهتمون بالأساليب التي تقوم باستكشاف دلالات البطاقات التصنيفية النصية (textual tags) بدلاً من ذلك (بما في ذلك علامات الهاشتاغ)، بالإضافة إلى الأبحاث في مجال نمذجة المستخدم المستندة إلى الدلالات في وسائل التواصل الاجتماعي.

حسب أنواع المعلومات الدلالية المستخدمة، يمكن تصنيف الأساليب كالتالي.

أكياس الكلمات ([336]) (Bag of words).

الكيانات التي يُزال عنها الغموض دلاليًّا: كيانات يذكرها المستخدم (مثال: [134، 337])، أو مأخوذة من مستند أطول موجود على شبكة الإنترنت (مثال: [134]).

الموضوعات: فئات موسوعة ويكيبيديا (مثال: [134، 338])، أو الموضوعات الكامنة (مثال [339]) أو تسلسلات بطاقات التصنيف الهرمية (مثال: [340]). من بين الحلول التي تُستخدم لنمذجة دلالات بطاقات التصنيف بصورة أكثر صراحة تفتيت بطاقات التصنيف وتحويلها إلى قاعدة معلومات WordNet ومن ثم استخدام مقاييس شبه دلالية تعتمد على WordNet لاشتقاق الصلة الدلالية لبطاقات فهارس المستخدمين (folksonomy) [341].

في العادة يتم تكملة ذلك بمعلومات اجتماعية ذات طابع كمي أكثر (عدد الارتباطات/ المتابعين لدى المستخدم مثلاً [231]) ومعلومات التفاعلات (على سبيل المثال: تكرار نشر المشاركات [232] ومعدل عدد المشاركات لكل موضوع [231]).

اكتشاف المعلومات الديموغرافية للمستخدمين

تعد مهمة اكتشاف المعلومات الديموغرافية للمستخدمين شديدة الأهمية في بناء نماذج المستخدمين باستخدام محتوى وسائل تواصل اجتماعي يتضمن شروحات دلالية. يوجد لدى كل مستخدم من مستخدمي موقع تويتر حساب خاص به يكشف بعض التفاصيل عن هويته. تكون حسابات المستخدمين شبه مهيكلة، وتتضمن حقلاً

خاصًّا بالمعلومات الذاتية واسم المستخدم الكامل وموقعه وصورة خاصة بالحساب والتوقيت الزمني ورابط الصفحة الرئيسة (معظم هذه المعلومات اختيارية وغالبًا ما تكون فارغة). يمكن الربط بين خصائص المستخدم ومحتوى مشاركاته، على سبيل المثال يمكن تحديد الموقع الجغرافي إلى حد ما من اللغة التي يستخدمها الشخص [342] أو الأحداث التي يُعلق عليها [343].

من بين تطبيقات أساليب معالجة اللغات الطبيعية اشتقاق المعلومات الديموغرافية الخاصة بالمستخدمين، عندما لا تكون متاحة بصورة جاهزة في حسابات وسائل التواصل الاجتماعي. من بين المهام التي يجري تناولها بصورة عامة تصنيف المستخدمين إلى ذكور أو إناث حسب نصوص تغريداتهم وحقول الوصف الخاصة بهم وأسمائهم، كما هو الحال مع دراسة [344]. في تلك الدراسة يعرض الباحثون دقة أعلى من معدلات الدقة البشرية مقارنة بأداء مجموعة من مضيفي الشروحات على موقع Mechanical Turk. كما جرى تطوير إطار عام لتصنيف المستخدمين بمقدوره أن يتعلم بصورة تلقائية كيفية اكتشاف الانتماءات السياسية والعرقية والمهتمين المتابعين لشركة معينة [345].

من الأبعاد المهمة الأخرى تحديد موقع مستخدمي تويتر بصورة تلقائية عبر تحليل محتوى مشاركاتهم وحساباتهم الشخصية[1]. تستخدم الأساليب عادة تقنيات معالجة اللغات الطبيعية لتحليل المحتوى النصي المقدم من قبل المستخدم واستنباط الموقع الجغرافي وفقًا للخصائص، مثل الإشارات التي تذكر أسماء المواقع المحلية [346] واستخدام اللهجات المحلية. في دراستي [342، 347] جرى اكتشاف مصطلحات ولغات خاصة بمناطق معينة قد تكون ذات صلة بالموقع الجغرافي للمستخدمين بصورة تلقائية. صممت دراسة [348] منهجية تصنيف تتضمن أيضًا إشارات محددة للأماكن القريبة من المستخدم. من مساوئ هذا الأسلوب أن شخصًا ما قد يكتب عن حدث عالمي مشهور لا يمت بصلة إلى موقعه الحقيقي. مثال آخر من مساوئ الأسلوب أن المستخدمين قد يتخذون خطوات مقصودة لإخفاء موقعهم الحقيقي عن طريق تغيير نمط مشاركاتهم أو تجنب الإشارة إلى المعالم المحلية.

1- http://openprovenance.org

استخدام الشروحات الدلالية لاشتقاق اهتمامات المستخدمين

من مجالات نمذجة المستخدمين المستندة إلى الدلالات التي تجري فيها الأبحاث بكثافة اشتقاق اهتمامات المستخدمين الضمنية باستخدام أساليب تمييز الكيانات، وكذلك نماذج الموضوعات. على سبيل المثال، استخدمت دراسة (أبيل وآخرون) [134] أدوات إضافة الشروحات الدلالية لاشتقاق حسابات المستخدمين بصورة آلية استنادًا إلى الكيانات والموضوعات. تجري نمذجة الحساب المستند إلى الكيانات والخاص بمستخدم معين في شكل مجموعة من الكيانات الموزونة، حيث يُحسب وزن كل كيان e بناءً إما على عدد تغريدات المستخدمين التي تذكر e أو بناءً على تكرار ورود الكيانات في التغريدات، بالإضافة إلى المقالات الإخبارية ذات الصلة (التي جرى تحديدها في خطوة ربط سابقة). تُعرّف الحسابات المستندة إلى الموضوعات بطريقة مشابهة، لكنها تمثل فئات موسوعة ويكيبيديا ذات المستوى المرتفع (كالرياضة والسياسة مثلاً). تُحدد الكيانات والموضوعات باستخدام برنامج OpenCalais (راجع القسم ٥-٤).

تستخدم دراسةُ (كابانيباثي وآخرون) [337] الشروحاتِ الدلالية بشكل مشابه لاشتقاق اهتمامات المستخدمين (الكيانات أو المفاهيم من DBpedia) التي توزن حسب قوتها (تُحسب بناءً على أساس تكرار الورود). كما تظهر كيف يمكن الدمج بين الاهتمامات بناءً على المعلومات المستمدة من مختلف وسائل التواصل الاجتماعي (لينكد إن وفيسبوك وتويتر). يجري جمع إعجابات فيسبوك والاهتمامات المذكورة صراحة في لينكد إن وفيسبوك مع معلومات الاهتمامات الضمنية المستمدة من التغريدات. يُستخدم نموذج Open Provenance Model [1] لتتبع أصل الاهتمامات.

اقتُرحت منهجية مشابهة مبنية على الكيانات والموضوعات لنمذجة اهتمامات المستخدمين من قبل مايكلسون وماكسكاسي [130] (تُدعى Twopics). تُعامل جميع الكلمات المكتوبة بالأحرف الكبيرة بخلاف كلمات التوقف التي ترد في التغريدات باعتبارها كيانات محتملة، ويجري البحث عنها في موسوعة ويكيبيديا (عناوين الصفحات ومحتوى المقالات). بعدها تأتي خطوة لإزالة الغموض حيث تحدد الكيان

1- http://leafletjs.com/

الموجود في موسوعة ويكيبيديا الذي يكون أفضل كيان مطابق للكيان المحتمل الموجود في التغريدة، في ضوء محتوى التغريدة الذي يُستخدم كسياق. لكل كيان يُزال الغموض عنه، يجري الحصول على شجرة فرعية لفئات موسوعة ويكيبيديا. في خطوة لاحقة تهدف لتحديد الموضوع، يجري تحليل جميع أشجار التصنيفات الفرعية لاكتشاف الفئات الأكثر تكرارًا، ومن ثم يتم تصنيفها كاهتمامات مستخدمين في ملفات المستخدمين المستندة إلى الموضوعات. يجادل المؤلفون أيضًا بأن مثل هذه الموضوعات الأكثر عمومية والتي يتم توليدها باستخدام تصنيف فئات موسوعة ويكيبيديا، تكون أنسب لعمليات التجميع والبحث عن المستخدمين من النماذج المستندة إلى المصطلحات المشتقة بواسطة أساليب كيس الكلمات (bag-of-words) أو LDA.

تسجيل سلوك المستخدم

كما سبق شرحه أعلاه، يعدُّ سلوم المستخدم عاملاً مهمًّا من العوامل المساعدة في فهم التفاعلات على وسائل التواصل الاجتماعي. في هذا القسم، نركّز في المقام الأول على المنهجيات التي تستخدم دلالات مشتقة آليًّا من أجل تصنيف سلوك المستخدم.

في حالة المنتديات الإلكترونية، جرى تصنيف أدوار سلوك المستخدم [349] التالية: نخبوي، ناخر، منضم للحوار، مبادر شعبوي، مشارك شعبوي، داعم، قليل الكلام ومُتجَاهَل. بالنسبة لأنظمة التصنيف الاجتماعي، قام الباحثون [350] بتقسيم المستخدمين حسب دافعهم للتصنيف إلى قسمين هما المصنفون والواصفون. في موقع تويتر، يجري رسم الدور الأكثر شيوعًا بناءً على محتوى التغريدات، ويُصنف المستخدمون إلى «meformers» (المغردين الذاتيين ويشكلون ٨٠٪ من المستخدمين) و«informers» (مغردي المعلومات ويشكلون ٢٠٪ من المستخدمين) [263].

من أجل تحديد أدوار سلوك المستخدم في المنتديات الإلكترونية بصورة آلية، قام (أنجليتو وآخرون) [231] بإنشاء هيكل قواعد بلغة سباركل (SPARQL) ترسم خريطة الخصائص الدلالية لتفاعل المستخدمين حسب مستوى السلوك (مرتفع ومتوسط ومنخفض). يجري إنشاء هذه المستويات بصورة ديناميكية من تفاعلات المستخدمين ويمكن تعديلها بمرور الوقت لمواكبة تطور المجتمعات الإلكترونية. كما

تجري نمذجة أدوار المستخدمين وسياقاتهم وتفاعلاتهم بصورة دلالية عبر أنطولوجيا سلوك المستخدم (راجع القسم ٨-٢) وتُستخدم لتوقع صحة منتدى إلكتروني معين.

لا تزال مشكلة تحديد خصائص سلوك مستخدمي تويتر حسب محتوى مشاركاتهم مجالاً لم يُستكشف بصورة وافية. قامت دراسة [237] بتوليد عبارات مفتاحية للمستخدمين بمساعدة وسيلة لنمذجة الموضوعات وأداة PageRank لترتيب الصفحات. وبالمثل تستخدم دراسة [234] مزيجًا يجمع بين تصفية أجزاء الكلام وأداة PageRank لاكتشاف بطاقات التصنيف الخاصة بالمستخدمين. ينبغي الملاحظة أيضًا أنه في حين قطعت دراسة [263] شوطًا مهمًّا نحو تصنيف سلوك المستخدم ونية التغريدات، إلا أن أسلوب الدراسة ليس آليًّا مع عدم وضوح ما إذا كان ممكنًا تحديد الفئات المماثلة بواسطة مصنف.

٩-٢-٢ النقاش

عند الحديث عن التغريدات، يمكن فصل اهتمامات المستخدمين المشتقة بصورة آلية إلى اهتمامات «عامة» (تستند إلى تغريدات المستخدم حول الموضوعات الرائجة) واهتمامات «خاصة بالمستخدم» (موضوعات تحمل طابعًا شخصيًّا بصفة كبرى كالعمل والهوايات والأصدقاء). هناك حاجة لإجراء مزيد من الدراسات حول التمييز بين الاهتمامات العامة (مثال: الأخبار الرائجة) والاهتمامات الخاصة بمستخدم معين (مثال: موضوع يتعلق بالعمل أو الهوايات أو إشاعة من صديق ...الخ). بعبارة أخرى، علينا تجاوز نطاق استخدام الشروحات الدلالية لتحديد ملفات المستخدمين بصورة آلية والانتقال نحو تحديد الأساس المنطقي والمصدر.

ترتبط الأشياء التي تعدُّ مهمة بالنسبة للمستخدم مع أدوار سلوك المستخدم (راجع القسم ٩-٢-١). ولذا يتطلب ذلك استخدام أساليب أكثر تعقيدًا لتحديد أدوار المستخدم بصورة آلية بناء على دلالات المشاركات، بالإضافة إلى الوسائل المستخدمة حاليًّا المبنية في المقام الأول على أنماط التفاعلات الكمية.

أخيرًا، هناك سؤال آخر يشكل تحديًا، وهو كيفية تجاوز نطاق النماذج المستندة إلى الاهتمامات والشبكات الاجتماعية القائمة على التفاعلات. على سبيل المثال، أظهرت دراسة (جينتايل وآخرون) [351] كيف يمكن استخلاص خبرات الأشخاص من رسائل البريد الإلكتروني التي يتبادلونها بينهم ومن ثم استخدامها لإنشاء ملفات مستخدمين تتسم بالديناميكية. بعد ذلك تجري المقارنة بين هذه الملفات من أجل اشتقاق شبكة مستخدمين تستند إلى الخبرات بدلاً من إنشاء شبكة مستندة إلى التفاعلات. يمكن توسيع نطاق منهجية كهذه وتكييفها لتناسب المدونات (مثال: من أجل استكشاف المدونات والتوصية بها)، وكذلك مشاركات تبادل البيانات المنشورة على موقعي تويتر ولينكد إن.

٩-٣ التصفية والتوصيات لمشاركات وسائل التواصل الاجتماعي

أدى الصعود غير المسبوق في حجم محتوى وسائل التواصل الاجتماعي وأهميته المتصورة إلى بدء شعور الأفراد بفيض المعلومات (information overload). في سياق استخدام الإنترنت، أشارت الدراسات التي تناولت فيض المعلومات أن وجود مستويات عالية من المعلومات يؤدي إلى عدم الفعالية، لأن «الشخص ليس بوسعه استيعاب جميع مُدخلات الاتصال والمعلومات» [352].

وعلى هذا النحو، قام باحثون بدراسة الأساليب المستندة إلى الدلالات لتصفية معلومات مشاركات وسائل التواصل الاجتماعي والتوصية بمحتواها. وبالنظر لكون الخطوط الزمنية في موقع فيسبوك ذات طابع خاص في معظمها، فقد ركز القسم الأكبر من الأعمال البحثية حتى الآن على موقع تويتر.

تشكل مشاركات وسائل التواصل الاجتماعي تحديًا من نوع خاص أمام وسائل التوصية بالمحتوى وتختلف عن الأنواع الأخرى من المستندات/ محتوى الويب، راجع دراسة [336]. بداية، ترتبط درجة صلة المحتوى بمدى حداثته، أي أن المحتوى لا يكون مثيرًا للاهتمام بعد مرور أيام على حدوثه. ثانيًا، يعدُّ المستخدمون مستهلكين ومنتجين نشطين للمحتوى الاجتماعي، كما أنهم مترابطون بشكل كبير بعضهم ببعض.

ثالثًا، يتعين على وسائل التوصية بالمحتوى تحقيق التوازن بين تصفية التشويش ودعم عنصر الصدفة/ اكتشاف المعرفة. أخيرًا، تختلف الاهتمامات والتفضيلات اختلافًا كبيرًا من مستخدم لآخر، وهذا يعتمد على حجم مشاركاتهم الشخصية والغرض الذي يستخدمون وسائل التواصل الاجتماعي من أجله وطريقة استخدامهم لها (راجع القسم ٩-٢-١ حول أدوار المستخدمين)، وسياق المستخدم (مثال: الأجهزة المحمولة مقابل الأجهزة اللوحية، العمل مقابل المنزل).

ركزت دراسة (تشين وآخرون) [336] و(أبيل وآخرون) [353] على تقديم توصيات لروابط URL لمستخدمي تويتر لكونها مهمة شائعة من مهام تبادل المعلومات. تعتمد منهجية دراسة (تشين وآخرون) على نموذج كيس-الكلمات (bag-of-words) الخاص باهتمامات المستخدمين، بناءً على تغريدات المستخدم، والموضوعات الرائجة دوليًّا والشبكة الاجتماعية الخاصة بالمستخدم. تجري نمذجة موضوعات روابط URL بصورة مشابهة كمتجه كلمة (word vector)، ويجري حساب توصيات التغريدات باستخدام شبه جيب التمام (cosine similarity).

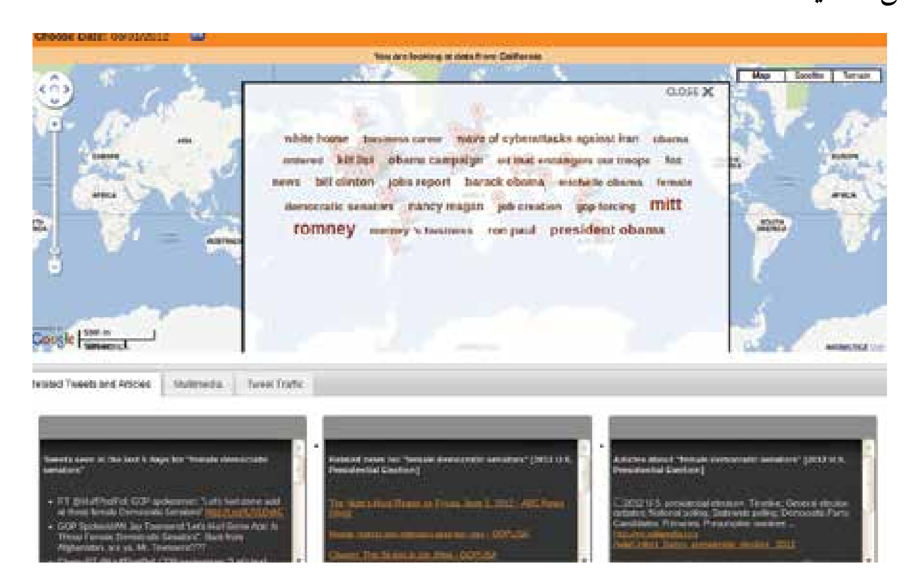
تقوم دراسة (أبيل وآخرون) [353] بتحسين هذه المنهجية باستخدام أدوات إضافة الشروحات الدلالية لاشتقاق نماذج اهتمامات المستخدمين المستندة إلى الدلالات (راجع القسم ٩-٢-١ لمزيد من التفاصيل). كما أنها تسجل قدرًا أكبر من الدلالات المتعمقة عن طريق تحليل دلالات علامات الهاشتاغ والردود وكذلك نمذجة الديناميكيات الزمنية لاهتمامات المستخدمين.

في دراسة حديثة أجراها (تشين وآخرون) [354] بتوسيع نطاق عمل الدراسة المذكورة أعلاه بالعمل من أجل التوصية بالنقاشات المهمة، أي موضوعات رسائل متعددة. يأتي الأساس المنطقي التي استندت عليه الدراسة من الاستخدام واسع الانتشار لموقعي فيسبوك وتويتر لإجراء النقاشات الاجتماعية [263]، إلى جانب الصعوبات التي تواجه المستخدمين في تتبع تلك المحادثات بمرور الوقت، ولا سيّما في موقع تويتر. يجري تصنيف النقاشات بناءً على طول النقاش وموضوعه (باستخدام نموذج كيس الكلمات كما ذكرنا أعلاه) وقوة الارتباط (تُعطى الأولوية للمحتوى

القادم من مستخدمين شديدي الترابط بعضهم ببعضهم). تترك الطبيعة السطحية لهذه المنهجية مساحة كبيرة لإجراء تحسينات من خلال استخدام الشروحات الدلالية وغيرها من أساليب معالجة اللغات الطبيعية التي ورد نقاشها في هذا الكتاب.

٩-٤ تصفح مشاركات وسائل التواصل الاجتماعي وعرضها بصيغة مرئية

يكمن التحدي الأكبر في تصفح الوسائل ذات المشاركات الضخمة وعرضها بصيغة مرئية في توفير نظرة شمولية عامة تكون في صيغة مجمّعة بدرجة مناسبة. في الغالب تكون واجهات القوائم المستندة إلى الطوابع الزمنية التي تعرض مشاركات كاملة يجري تحديثها بصورة متواصلة (مثال: واجهة الويب المستندة إلى الخط الزمني في موقع تويتر) غير عملية، ولا سيّما في تحليل الأحداث ذات الأحجام الكبيرة والتي تحدث بصورة متقطعة. على سبيل المثال، خلال حفل الزفاف الملكي الذي جرى في عام 2011، تجاوز عدد التغريدات حاجز المليون. وبالمثل تكون مراقبة الأحداث التي تستمر لمدة طويلة، كحملات الانتخابات الرئاسية، في مختلف الوسائل والمواقع الجغرافية، بالدرجة نفسها من التعقيد.

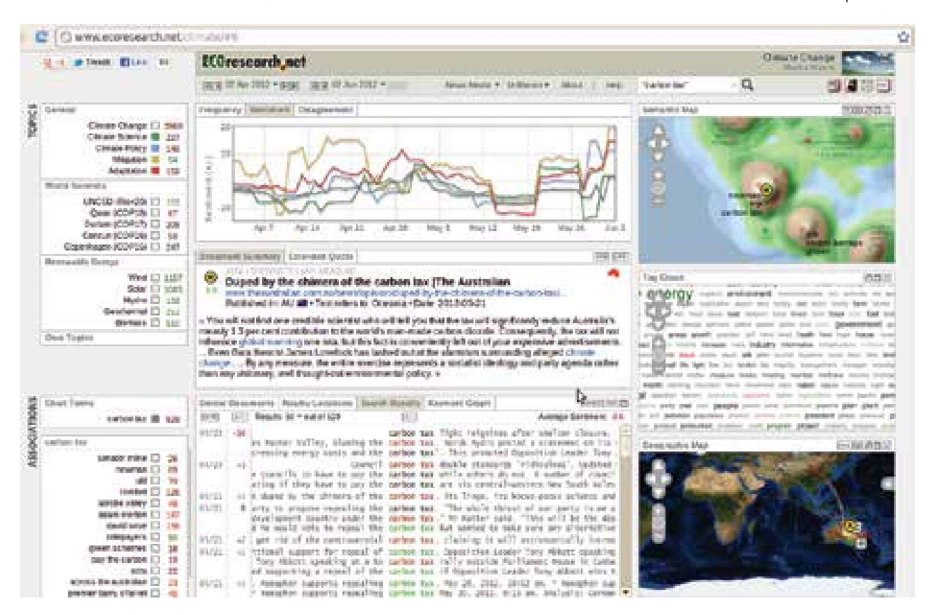


الشكل ٩-٩: منصة Twitris **لمراقبة أحداث وسائل التواصل الاجتماعي** (http: / /twitris.knoesis.org).

تعدُّ سحابات الكلمات (word clouds) من أبسط التصويرات الرسومية وأكثرها استخدامًا. تستخدم هذه السحابات عمومًا مصطلحات مكونة من كلمة واحدة، وهو ما قد يصعب تفسيره من دون وجود سياق إضافي. استُخدمت سحابات الكلمات لمساعدة المستخدمين في تصفح مشاركات وسائل التواصل الاجتماعي، بما في ذلك محتوى المدونات [355] والتغريدات [261، 356]. على سبيل المثال، استخدم (فيلان وآخرون) [357] سحابات الكلمات لعرض نتائج نظام توصية يستند إلى تويتر. بدوره يستخدم نظام إيدي [358] سحابات الموضوعات، حيث يعرض موضوعات أكثر شمولية في سلسلة تغريدات المستخدم. يجري الجمع بين هذه السحابات وقوائم الموضوعات التي تعرض الأشخاص الذين كتبوا تغريدات عن الموضوعات، وكذلك مجموعة من التغريدات المثيرة للاهتمام لأعلى الموضوعات تصنيفًا. يشتق نظام Twitris (راجع الشكل ٩-٩) عددًا أكبر من العبارات السياقية الأكثر تفصيلاً باستخدام 3-grams بدلاً من uni-grams [261]. في الآونة الأخيرة، جرى توسيع نطاق المفهوم ليشمل سحابات الصور [254].

يكمن العيب الرئيس للتصويرات الرسومية المستندة إلى السحابات في طبيعتها الثابتة. لذا فإنها غالبًا ما تُدمج مع الخطوط الزمنية التي تظهر تكرارات الكلمات المفتاحية/ الموضوعات بمرور الوقت [260، 273، 358، 359]، بالإضافة إلى أساليب اكتشاف الارتفاعات غير العادية في مستويات الشعبية [355]. تستخدم دراسة [269] خطًّا زمنيًّا متزامنًا مع نص بث تلفزيوني سياسي، ما يتيح الانتقال إلى النقاط الرئيسة في الفيديو الخاص بالحادثة، وعرض التغريدات المنشورة في تلك الفترة الزمنية. كما يجري عرض الشعور العام في خط زمني في كل نقطة في الفيديو، وذلك باستخدام شرائح ملونة بسيطة. وبالمثل يستخدم نظام TwitInfo (راجع الشكل ٩-١١ [262]) خطًّا زمنيًّا لعرض نشاط التغريدات أثناء وقوع أحداث حقيقية في العالم (مثال: لعبة كرة قدم) إلى جانب عدد من التغريدات النموذجية المرمزة بالألوان للإشارة إلى المشاعر. تكون بعض هذه التصويرات الرسومية ذات طابع ديناميكيي، أي أنه يجري تحديثها مع وصول محتوى جديد (مثال: تيارات الموضوعات [254]، أشرطة الكلمات المفتاحية

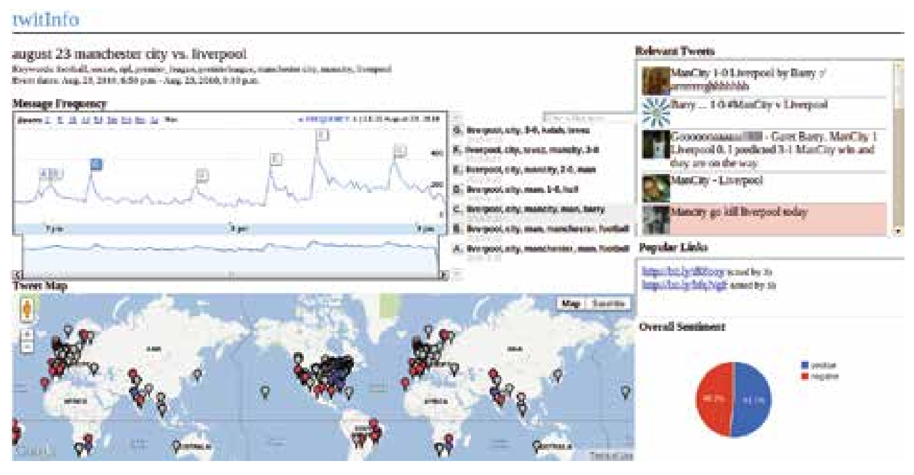
المنحدرة [273] مناظر المعلومات الديناميكية [273])، أو أشرطة العنوانات التي تقارن التغريدات بجانب معايير مختلفة (في هذه الحالة، انقسم ناشرو التغريدات حسب دعمهم لحملة مغادرة/ بقاء المملكة المتحدة في الاتحاد الأوروبي، الشكل ٩-١٣).



الشكل ٩-١٠: مراقبة وسائل الإعلام في منصة التغير المناخي
.(http: / /www.ecoresearch.net /climate)

علاوة على ذلك، تحاول بعض التصويرات الرسومية تسجيل الترابط الدلالي بين الموضوعات في مشاركات وسائل التواصل الاجتماعي. على سبيل المثال، يقوم نظام BlogScope [355] بحساب الارتباطات بين الكلمات المفتاحية عن طريق تقدير المعلومات المتبادلة لِزَوج من الكلمات المفتاحية باستخدام عينة عشوائية من المستندات. هناك مثال آخر وهو التصوير الرسومي لمشهد المعلومات الذي يعرض الشبه بين الموضوعات من خلال القُرب المكاني (spatial proximity) (راجع الشكل ٩-١٠). يمكن أيضًا عرض العلاقة بين الموضوعات والمستندات عن طريق التصويرات الرسومية الموجهة بالقوة والمستندة إلى الرسوم البيانية [360]. أخيرًا، تقترح دراسة (آرشامبو وآخرون) [361] سحابات بطاقات تصنيف متعددة المستويات من أجل تسجيل العلاقات الهرمية.

هناك بعد مهم آخر من أبعاد المحتوى المُنتج من قبل المستخدم، وهو مكان المنشأ. على سبيل المثال، يجري إضافة بطاقات تصنيف جغرافية تحمل معلومات خطوط العرض/ الطول إلى التغريدات، في حين تحدد الكثير من ملفات المستخدمين على موقعي فيسبوك وتويتر وكذلك المدونات مكان المستخدم. وبناءً على ذلك، جرى استكشاف التصويرات الرسومية المستندة إلى الخرائط [261، 262، 273، 262] (انظر أيضًا الرسم ٩-١٠ والرسم ٩-١١). على سبيل المثال، يسمح نظام Twitris [261] للمستخدمين اختيار دولة معينة من خرائط جوجل ويعرض الموضوعات التي يجري نقاشها في وسائل التواصل الاجتماعي من هذه الدولة فقط. يعرض الشكل 9-9 نظام Twitris أثناء مراقبة الانتخابات التي جرت في عام ٢٠١٢ في الولايات المتحدة، حيث اخترنا مشاهدة الموضوعات ذات الصلة التي يجري نقاشها في وسائل التواصل الاجتماعي والتي يكون منشؤها في ولاية كاليفورنيا. عند الضغط على موضوع «أعضاء مجلس الشيوخ من النساء»، يجري عرض التغريدات والأخبار ومقالات موسوعة ويكيبيديا ذات الصلة. للمقارنة، يعرض الشكل ٩-١٢ الموضوعات التي تحظى بأكبر قدر من النقاش المتعلقة بالانتخابات والتي استُخرجت من مشاركات على وسائل التواصل الاجتماعي يعود أصلها إلى بريطانيا العظمى. وفي حين يوجد تداخل كبير بين الموقعين الجغرافيين، لكن الاختلافات تبدو واضحة أيضًا.



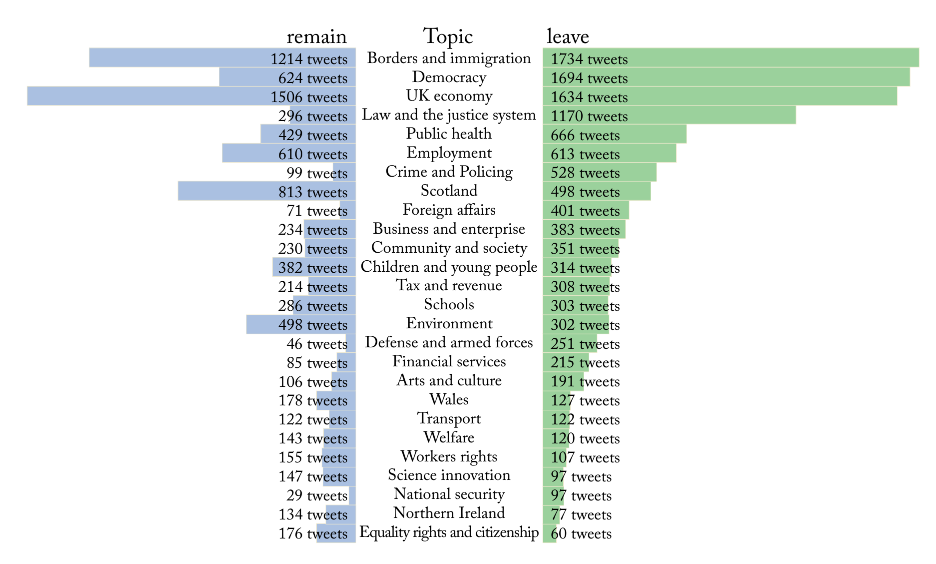
الشكل ٩-١١: نظام TwitInfo **متعقبًا إحدى مباريات كرة القدم** (http: //twitinfo.csail.mit.edu/).

من الممكن تجميع التغريدات وعرضها بصيغة رسومية بناءً على موقع وجود ناشر التغريدة، بمعنى التحقيق في التباينات الجغرافية بين الموضوعات المذكورة. يظهر المثال المعروض أدناه تصويرات رسومية تستند إلى نظام Mímir وتعرض الموضوعات التي يجري الحديث عنها أكثر في مختلف أجزاء البلاد، بناءً على تجميع التغريدات المنشورة من قبل مرشحي الانتخابات البريطانية حسب تصنيف أقاليم نظام NUTS لتصنيف أقاليم دول الاتحاد الأوربي. يتضمن ذلك إصدار سلسلة من استفسارات Mímir عن التغريدات لكل موضوع، من أجل معرفة عدد التغريدات التي تذكر كل موضوع والتي كتبها كل عضو في البرلمان يمثل كل إقليم. لا يتم التعبير عن المعلومات المتعلقة بالإقليم الذي يمثله عضو البرلمان في التغريدة نفسها، لكنها تستخدم قاعدتنا المعرفية بمرحلتين: الأولى هي إيجاد الدائرة التي يمثلها عضو البرلمان، ومن ثمّ مطابقة الدائرة مع الإقليم المناسب وفقًا لتصنيف NUTS. يبيّن الشكل ٩-١٤ خريطة كوروبليث (choropleth) تعرض توزيع تغريدات أعضاء البرلمان التي تناقش اقتصاد المملكة المتحدة (وهو الموضوع الأكثر تكرارًا) في التغريدات المنشورة خلال الانتخابات البريطانية العامة التي جرت في عام ٢٠١٥ والتي جرى جمعها في الأسبوع الذي كانت بدايته ٢ مارس ٢٠١٥. تعدُّ الخريطة تصويرًا مرئيًّا ديناميكيًّا يعتمد على مكتبة Leaflet [1]، ويقوم نظام Mímir بعرض النتائج المجمّعة للاستفسار لكل موضوع وإقليم NUTS1. يوجد في choropleth قائمة منسدلة يمكن للمستخدم أن يختار منها الموضوع الذي يهمه، وهو ما يؤدي إلى إعادة رسم الخريطة وفقًا لذلك. تتوفر نسخ تجريبية وchoropleth وشجرة خريطة تفاعلية في مجموعة البيانات هذه، وكذلك أمثلة على سحابة الموضوعات وتصوير رسومي للمشاعر، بصورة يمكن الاطلاع عليها من خلال هذا الرابط
http: / /www.nesta.org.uk /blog /4-visualizationsuk-general-election.

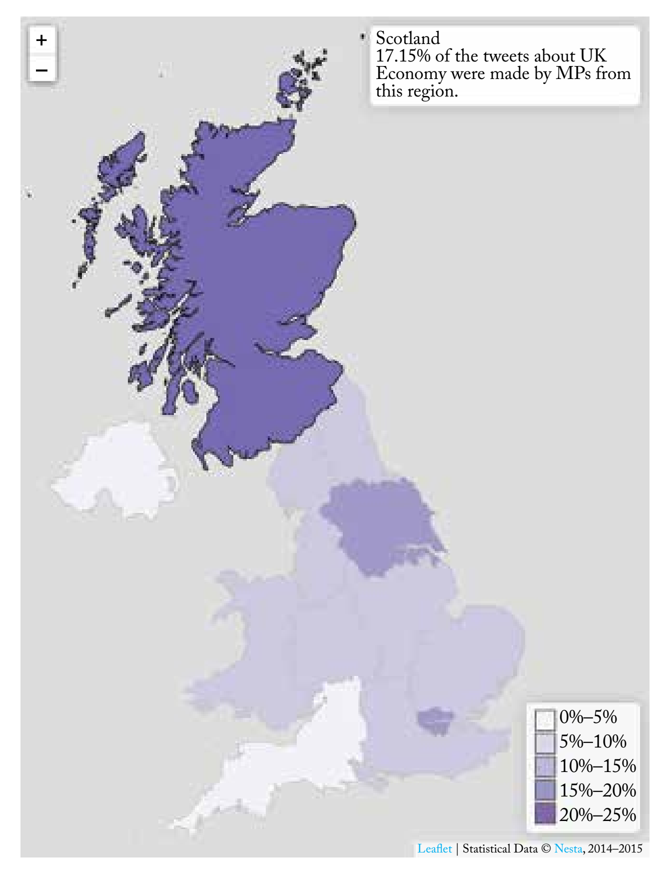
1- http://cloud.gate.ac.uk



الشكل ٩-١٢: الموضوعات المختلفة المستخرجة بواسطة نظام Twitris فيما يتعلق ببريطانيا العظمى.



الشكل ٩-١٣: أشرطة الموضوعات التي تقارن بين التغريدات المنشورة حول تلك الموضوعات من قبل داعمي حملتي استفتاء مغادرة الاتحاد الأوروبي أو البقاء فيه.



الشكل ٩-١٤: خريطة كوروبليث (Choropleth) تبيّن توزيع التغريدات التي تتناول الاقتصاد.

كما تظهر الآراء والمشاعر بصورة متكررة في واجهات التحليلات المرئية. على سبيل المثال، يجمع نظام Media Watch (الشكل ٩-١٠ [273]) بين سحابات الكلمات وقطبية المشاعر المجمعة، حيث تُلوّن كل كلمة بإحدى درجات اللون الأحمر (المشاعر السلبية بالدرجة الأولى) أو اللون الأخضر (المشاعر الإيجابية بالدرجة الأولى) أو اللون الأسود (المشاعر المحايدة). كما يجري تلوين مقتطفات نتائج البحث ومصطلحات التصفح المتعددة بألوان تشير إلى المشاعر. كما جمع آخرون بين الترميز بالألوان استنادًا إلى المشاعر والخطوط الزمنية للأحداث [359] وقوائم التغريدات (الشكل ٩-١١ [262]) وخرائط المزاج [359]. في العادة يجري عرض المشاعر المجمّعة باستخدام الرسوم البيانية الدائرية [260]، وفي حالة نظام TwitInfo، يجري تطبيع الإحصاءات

الإجمالية لغرض الاستدعاء (الرسم ٩-١١ [262]).

كما قام الباحثون بالتحقيق تحديدًا في مشكلة تصفح محادثات وسائل التواصل الاجتماعي المتعلقة بالأحداث العالمية وتصويرها رسوميًّا، مثل الأحدث التي يجري بثها على الهواء [356] ومباريات كرة القدم (الرسم ٩-١١ [262]) والمؤتمرات [254] وأحداث الأخبار [359، 362]. هناك عنصر مهم، وهو القدرة على تحديد الأحداث الفرعية وجمعها مع الخطوط الزمنية والخرائط والتصويرات الرسومية المستندة إلى الموضوعات.

جرى أيضًا تصميم تصويرات رسومية أخرى للاستفادة من جهة من كون مشاركات وسائل التواصل الاجتماعي محتوى ينتجه المستخدمون، وطابعها الاجتماعي من جهة أخرى. على سبيل المثال يرسم نظام PeopleSpiral للتصوير الرسومي [254] مستخدمي تويتر الذين شاركوا في أحد الموضوعات (مثال: نشر التغريدات باستخدام علامة هاشتاغ معينة) المنتشرة بصورة متصاعدة، بداية بالمستخدمين الأكثر نشاطًا و«أصالة». يجري قياس أصالة المستخدم كنسبة بين عدد التغريدات المكتوبة من قبل المستخدم مقارنة بالتغريدات المعاد نشرها. بدلاً من ذلك يقوم نظام OpinionSpace [363] بتصوير المستخدمين رسوميًّا في مساحة ثنائية الأبعاد، بناءً على الآراء التي عبروا عنها في مجموعة معينة من الموضوعات. تظهر كل نقطة في التصوير الرسومي أحد المستخدمين وتعليقه، لذا كلما كانت النقطتان بعضهما أقرب لبعض كانت آراء المستخدمين أكثر شبهًا بعضها ببعض. غير أن التصوير الرسومي المستند إلى النقاط بصورة محضة ثبت أنه صعب التفسير من قبل بعض المستخدمين، وذلك لأنهم غير قادرين على رؤية المحتوى النصي حتى يقوموا بالضغط على إحدى النقاط. بدلاً من ذلك، يقوم نظام ThemeCrowds [361] باشتقاق تجميعات هرمية لمستخدمي تويتر عبر تجميع الكتل (agglomerative clustering) ويقدِّم ملخصًا للتغريدة التي يجري إنتاجها من قبل هذه الكتلة، عن طريق سحابات بطاقات تصنيف متعددة المستويات (المستوحاة من تصوير شجرة الخريطة الرسومية). تُعرض أحجام التغريدات بمرور الوقت بأسلوب مشابه للخط الزمني، وهو ما يسمح أيضًا باختيار الفترة الزمنية.

٩-٥ النقاش والأعمال المستقبلية

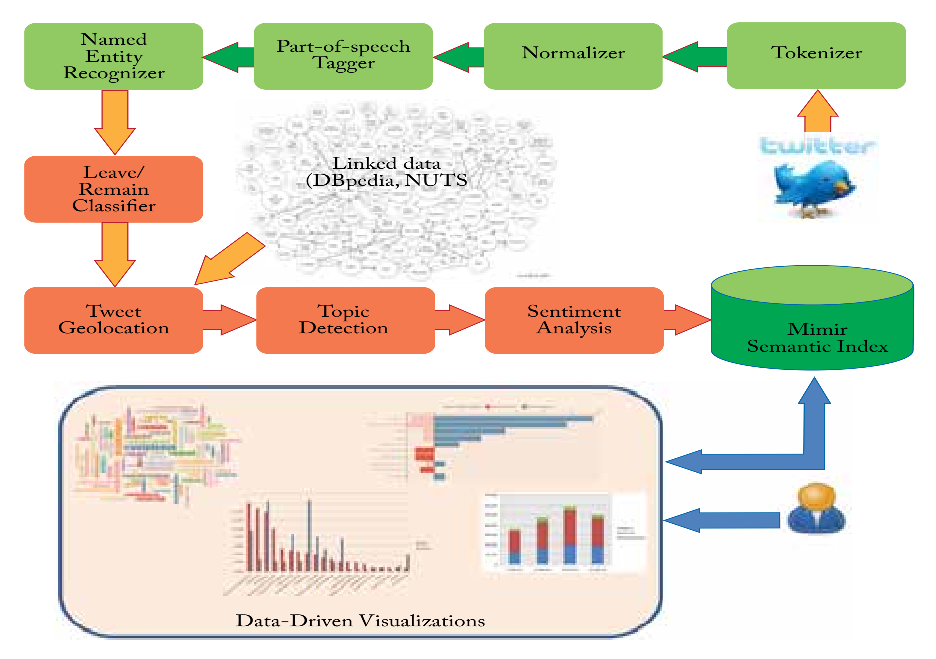
أغلب الأبحاث والتوصيات وأساليب التصوير الرسومي تميل إلى استخدام معلومات سطحية نصية ومعلومات مستندة إلى التكرار. على سبيل المثال، أظهرت مقارنة بين نمذجة الموضوعات الموزونة وفق تكرار المصطلح-عكس تكرار المستند (TF-IDF) ونمذجة LDA للموضوعات أن الأولى أكثر تفوقًا [238، 354]. تقترح دراسة [354] أنه يمكن تحسين هذه النماذج بشكل أكبر عن طريق الدمج بين المعلومات الدلالية. في حالة التوصيات التي تحمل الطابع الشخصي، يمكن تحسين هذه النماذج من خلال إضافة أدوار سلوك المستخدم، وهو ما يستغل الدلالات الكامنة ومعلومات المستخدم الضمنية استغلالاً أفضل، ويؤدي إلى دمج البُعد الزمني في الخوارزميات المقترحة.

يمكن أيضًا تحسين واجهات التصفح والتصوير الرسومي عن طريق أخذ المعرفة الدلالية الإضافية عن الكيانات المذكورة في المشاركات في الاعتبار. على سبيل المثال، عندما تُضاف الشروحات إلى الموضوعات بواسطة روابط URI تؤدي إلى مصادر LOD، مثل DBpedia، يمكن أن تدعم الأنطولوجيا الكامنة تصويرات رسومية ذات تسلسل هرمي، بما في ذلك العلاقات الدلالية. إضافة إلى ذلك، يمكن إثراء عملية استكشاف مشاركات وسائل التواصل الاجتماعي من خلال تصويرات رسومية مبنية على الموضوعات والكيانات والوقت باستخدام واجهات البحث المتعدد والاستعلام الدلالي التي تعتمد على الأنطولوجيات. من الأمثلة على ذلك منصة KIM الدلالية الموجهة نحو مجموعات المستندات التي تكون ثابتة إلى حد بعيد [317].

تعدُّ قابلية الخوارزميات للتوسيع ومدى كفاءتها من العناصر ذات الأهمية الخاصة، وذلك بسبب سعة نطاق مشاركات وسائل التواصل الاجتماعي وطبيعتها الديناميكية. على سبيل المثال، تستغرق منصة Topic Stream التفاعلية 45 ثانية لحساب مليون تغريدة و325000 مستخدم مشارك، وهو ما يعدُّ طويلاً جدًّا لمعظم سيناريوهات الاستخدام [254]. وبالمثل يعدُّ حساب الارتباطات بين الكلمات عن طريق المعلومات النقطية التبادلية (pointwise mutual information) باهظ الثمن من الناحية الحسابية فيما يتعلق بالمدونات ذات الحجم الكبير [355]. هناك حل يتم استخدامه بصورة

متكررة، وهو وضع نافذة متحركة فوق النص (مثال: بين أسبوع واحد وسنة واحدة) ومن ثم يتم تحديد حجم المحتوى المستخدم لـ IDF وغير ذلك من العمليات الحسابية.

معظم الأنظمة والمنهجيات التي تم استعراضها هنا ليست قابلة للتوسيع أو التكييف بسهولة مع مشكلة جديدة أو مع تصوير رسومي جديد أو مع قدرات إضافة الشروحات الدلالية ذات النطاق الواسع. تكمن فائدة الأدوات ذات المصدر المفتوح المعتمدة على نظام Gàte والتي تستخدم للبحث والتصوير الرسومي الدلالي (نظام Mímir ونظام Prospector) ومنظومة نظام GATE للتحليلات التفاعلية في أنها ذات مصدر مفتوح قابل للتوسيع والتمديد. خلال تطبيق هذه الأدوات مؤخرًا في تحليل تغريدات استفتاء خروج بريطانيا من الاتحاد الأوروبي (أي محل البريكست، راجع الشكل ٩-١٥)، كان متوسط عدد التغريدات اليومية نحو ٥٠٠,٠٠٠ تغريدة يوميًّا، وكانت ذروة عدد التغريدات مليوني تغريدة في يوم التصويت. هذا الأمر تطلَّب توفر مكونات عالية الأداء لإجراء التحليلات الدلالية والفهرسة والبحث والتصوير الرسومي، وصُممت تلك المكونات لتحليل ما يصل إلى ١٠٠ تغريدة في الثانية الواحدة.



الرسم ٩-١٥: بنية نظام التحليل الدلالي والبحث والتصوير الرسومي لحملة الـ«بريكست».

لإجراء التحليلات، نستخدم نظام TwitIE التابع لمنصة GATE [248]، ويتكون النظام من أداة تجزئة الوحدات اللغوية وأداة إعادة النص للشكل القياسي وأداة تصنيف أقسام الكلام وأداة تمييز كيانات الأسماء. بعد ذلك، أضفنا أداة لتصنيف التغريدات إلى تغريدات مغادرة الاتحاد الأوروبي وتغريدات البقاء فيه، وذلك لتحديد عيّنة موثوق بها من التغريدات ذات المواقف غير الملتبسة. بعدها يأتي دور مكون تحديد الموقع الجغرافي للتغريدة، حيث يستخدم بيانات خطوط الطول/ العرض والإقليم وموقع المستخدم من أجل تحديد الموقع الجغرافي للتغريدات داخل أقاليم نظام UK NUTS2 لتصنيف أقاليم المملكة المتحدة. جرى اكتشاف الموضوعات الرئيسة التي نوقشت في التغريدات (قد تحمل كل تغريدة أكثر من موضوع واحد)، وبعدها يأتي دور تحليل المشاعر المتمحور حول الموضوعات. كانت الفائدة الرئيسة في استخدام عدد كبير من مكونات إضافة الشروحات الدلالية المتوفرة مسبقًا في أن تطوير التطبيق استغرق وقتًا قصيرًا للغاية.

تدعم عمليات البحث والتصوير الرسومي المستندة إلى نظام Mímir استكشاف مجموعات بيانات كبيرة تتألف من أكثر من ٦٤ مليون تغريدة بصورة فعالة. تحتوي استعلامات Mímir الاعتيادية قيودًا من قبيل الطابع الزمني (يُحوّل إلى توقيت جرينيتش) ونوع التغريدة (تغريدة أصلية أو رد على تغريدة أخرى أو إعادة نشر تغريدة أخرى) ونية التصويت (المغادرة/ البقاء) وذكر مستخدم/ هاشتاغ/ موضوع معين، وكتابة التغريدة من قبل مستخدم محدد، واحتواء التغريدة على علامة هاشتاغ معينة أو موضوع محدد (مثال: جميع التغريدات التي تناقش الضرائب). يوجد أعلاه تصويرات رسومية تستند إلى الشروحات نقدمها كأمثلة. تتميز جميع هذه التصويرات بأنها تفاعلية، حيث يستطيع المستخدم الضغط على عنصر معين (مثال: شريط موضوعات أو إقليم NUTS) ورؤية جميع التغريدات التي تدعم هذا العنصر المحدد من عناصر التصوير الرسومي المجمّعة بصورة فورية. ومع أنها ما زالت في مرحلة التطوير، إلا أن هذه المنهجية المفتوحة المصدر والخاصة بعمليات البحث والتصوير الرسومي ذات النطاق الواسع قد أثبتت قدرتها على توفير مزايا عديدة من حيث تقليل الوقت المستغرق في التطوير وفي مستوى الفعالية وقدرتها على توفير تصويرات رسومية متعددة.

ختامًا، أثبت تصميم واجهات فعالة للبحث الدلالي والتصفح والتصويرات الرسومية للمشاركات ذات الحجم الكبير والسرعة المرتفعة أنه يطرح تحديًا من نوع خاص.

تشمل بعض المشكلات التي تحتاج مزيدا من البحث والدراسة ما يلي:

- تصميم تصويرات رسومية بديهية وذات معنى قادرة على أن تعبر بصورة بديهية الدلالات المعقدة ذات الأبعاد المتعددة للمحتوى المُنتج من قبل المستخدم، (على سبيل المثال الموضوعات والكيانات والأحداث والمعلومات الديموغرافية الخاصة بالمستخدم (بما في ذلك المواقع الجغرافية والمشاعر والشبكات الاجتماعية).
- عرض التغييرات التي تحدث بمرور الوقت بصيغة رسومية.
- دعم المستويات المختلفة من التجزئة التفصيلية (granularity) على مستوى المحتوى الدلالي ومجموعات المستخدمين والنوافذ الزمنية.
- السماح باستكشاف تفاعلي لحظي.
- التكامل مع البحث للسماح للمستخدمين باختيار جزء فرعي من المحتوى ذي الصلة.
- إزاحة الستار عن الطابع النقاشي/ الموضوعي للمحادثات الدائرة على وسائل التواصل الاجتماعي، ومعالجة المشكلات المتعلقة بقابلية التوسيع والكفاءة.

الفصل العاشر
الخاتمة

نختتم هذا الكتاب بملخص للنقاط الرئيسة وبعض الملحوظات العامة حول استخدام معالجة اللغات الطبيعية في تطبيقات الويب الدلالي، وبعض الأفكار عن الاتجاهات المستقبلية.

١٠-١ ملخص

كان هدف هذا الكتاب تقديم بعض المفاهيم والأساليب والأدوات الأساسية في معالجة اللغات الطبيعية وتحليل النصوص وعرضها أمام باحثي الويب الدلالي، وشرح الأسباب التي تجعلها ضرورية لتكوين فهم واضح ليس لجعل أساليب معالجة اللغات الطبيعية مفيدة فحسب، بل أيضًا لفهم أوجه القصور فيها. شرحنا هذه الأساليب في مختلف فصول الكتاب مع عرض أمثلة للأدوات الشائعة ذات المصدر المفتوح التي يمكن استخدامها، وناقشنا المشكلات المتعلقة بدمج تلك الأدوات المعتمدة، وإعطاء فكرة معينة عن الأداء المتوقع.

جرى تخصيص الجزء الأول من هذا الكتاب لشرح المفاهيم الرئيسة التي تشكل الأساس لعملية معالجة اللغات الطبيعية، وذلك من أجل التمهيد لمهام أكثر تعقيدًا في المراحل التالية من الكتاب. حرصنا كثيرًا على اتباع منهجية «خط الأنابيب» (pipeline) المتبعة في بناء التطبيقات المعتمدة على معالجة اللغات الطبيعية، بداية بالمهام ذات المستوى المنخفض مثل مهام معالجة اللغات الطبيعية الأساسية، ثم الانتقال إلى مهام أكثر تعقيدًا مثل مهام إيجاد العلاقات وتطوير الأنطولوجيات وتعدين الآراء. كما وضعنا في الاعتبار أنواعًا مختلفة من المهام والتطبيقات، مثل تحليل وسائل التواصل الاجتماعي وأنواع التكييفات المحددة المطلوبة لإجراء تلك المهام، بالإضافة إلى كيفية استخدام جميع هذه الأدوات لإنشاء تطبيقات أكثر تعقيدًا كالتطبيقات المعززة دلاليًّا لاسترجاع المعلومات وعرضها في صيغة مرئية.

في نهاية المطاف، يفترض أن يخرج القارئ بعد قراءة هذا الكتاب بفهم المبادئ الرئيسة لمعالجة اللغات الطبيعية ودورها في الويب الدلالي، ولديه القدرة على اختيار تقنيات معالجة اللغات الطبيعية التي يمكن استخدامها لتعزيز تطبيقات الويب الدلالي

الخاصة به. هناك بالطبع الكثير من الموضوعات والأدوات التي لم نناقشها هنا، ولكن أشرنا إلى مراجع وأماكن أخرى يمكن العثور فيها على شروحات أكثر تفصيلاً. يحاول هذا الكتاب أن يجمع في مكان واحد بعض المواد التي تعد الأكثر صلة لتحقيق هذه الغايات.

١٠-٢ الاتجاهات المستقبلية

في حين تشكل الأساليب الجوهرية لمعالجة اللغات الطبيعية الأساس الذي يقوم عليه الكثير من مهام معالجة اللغات الطبيعية، مثلما لاحظنا في مختلف أقسام هذا الكتاب، إلا أنه لا تزال هناك العديد من التحديات التي ينبغي مواجهتها عند اعتماد أساليب وأدوات معالجة اللغات الطبيعية وتكييفها لتتلاءم مع الأشكال الجديدة للبيانات والأنواع الجديدة من التطبيقات التي تظهر باستمرار. في هذا القسم، نناقش بعضًا من الاتجاهات المهمة التي ينبغي أن تمضي نحوها أبحاث معالجة اللغات الطبيعية من أجل مواكبة التطورات التكنولوجية.

١٠-٢-١ التجميع متعدد الوسائط والتعدد اللغوي

جرى تطوير غالبية الأساليب المشمولة في هذا الكتاب وتقييمها على نوع واحد فقط من أنواع الوسائط (مثال: النصوص الإخبارية أو تويتر أو مشاركات المدونات). غير أن العديد من التطبيقات الحالية يتطلب دمج أنواعٍ مختلفة من النصوص، على سبيل المثال ربط التغريدات بالمقالات والمدونات الإخبارية. علاوة على ذلك، يمكن أن يتجاوز الربط بين الأنواع المختلفة من الوسائط هذا النطاق، وهذه قضية مهمة ما زالت مفتوحة، وذلك بسبب كون المستخدمين يستخدمون أكثر من منصة واحدة من منصات وسائل التواصل الاجتماعي، وغالبًا ما يكون ذلك لأسباب مختلفة (مثال: لأغراض الاستخدام الشخصي مقارنة بأغراض الاستخدام المهني). إضافة إلى ذلك، وفي ضوء تحول أسلوب حياة الناس إلى أسلوب رقمي على نحو مطّرد، سيقدم هذا العمل إجابة جزئية تسهم في التغلب على التحدي الذي تمثله عملية الربط بين مجموعاتنا الشخصية (مثال: رسائل البريد الإلكتروني، الصور) مع هوياتنا على وسائل التواصل الاجتماعي.

يكمن التحدي في بناء نماذج حسابية لدمج محتوى الوسائط المتعددة وتحليلها وعرضها في صيغة مرئية، وتضمينها في خوارزميات قادرة على التعامل مع تدفقات وسائل التواصل الاجتماعي ذات المنصات المتعددة التي تتسم بكونها ذات أعداد كبيرة وذات طبيعة متناقضة ومتعددة الأغراض. على سبيل المثال، هناك حاجة لإجراء المزيد من الأعمال على خوارزميات تجميع محتوى الوسائط المتعددة ورصد الهويات على الوسائط المتعددة ونمذجة التناقضات بين المصادر المختلفة، واستنباط التغيرات التي تطرأ على الاهتمامات والسلوكيات مع مرور الوقت.

هناك تحدٍّ كبير آخر ذو صلة، وهو تحدي التعددية اللغوية، فمعظم الأساليب المشمولة في هذا الكتاب جرى تطويرها واختبارها باستخدام محتوى مكتوب باللغة الإنجليزية فقط، لأنها عادة ما تكون أول باب تطرقه الأساليب التكنولوجية والتطبيقات الجديدة. غير أنه ينبغي لنا ألاَّ نتغاضى عن أهمية تكييف هذه الأدوات لتتلاءم مع اللغات الأخرى و/ أو تمكينها من التعامل مع لغات متعددة في آن واحد. وكما ناقشنا في القسم ٨-٣-٧، يجري اتخاذ بعض الخطوات الأولية عبر توفير معاجم متعددة اللغات، مثل Wiktionary [289] وUBY [290]، والأنطولوجيات القائمة على أسس لغوية [291]. كما ركزت الأبحاث الأخرى على توسيع نطاق الموارد اللغوية المتوفرة للغات التي تجري دراستها بصورة أقل، وذلك عبر ما يُعرف بالتعهيد الجماعي (crowdsourcing) وهي الاستعانة بالجمهور من أجل الحصول على البيانات أو المعلومات. على وجه الخصوص، برزت خدمة «أمازون ميكانيكال تورك» (Amazon Mechanical Turk) كأداة مهمة، وذلك لسهولة إنشاء مشاريع التعهيد الجماعي فيها، إلى جانب كونها تسمح بـ»الوصول إلى أسواق أجنبية يوجد فيها أشخاص يتحدثون الكثير من اللغات النادرة» [364]. تكون هذه الخدمة مفيدة بصفة خاصة للباحثين الذين يعملون على اللغات المنخفضة الموارد كالعربية [365] والأوردية [364] وغيرهما [366-368]. تبيّن دراسة إيرفين وكليمينتييف [368] على سبيل المثال أنه يمكن إنشاء معاجم تجمع بين اللغة الإنجليزية و٣٧ من أصل الـ٤٢ لغة منخفضة الموارد التي شملتها اختبارات الدراسة. وبالمثل تقوم دراسة (فايكسلبراون وآخرون) [369] بإنشاء معاجم مشاعر ذات نطاق محدد

عبر التعهيد الجماعي بلغات عدة، وذلك عبر ألعاب هادفة. من الجوانب ذات الصلة تصميم مشاريع التعهيد الجماعي لكي يسهل استخدامها مرة أخرى بلغات متعددة، على سبيل المثال [368، 370] بالنسبة لخدمة «أمازون ميكانيكال تورك» و[371، 372] بالنسبة للألعاب الهادفة. هناك أيضًا مسألة متصلة تتعلق بالمكانز ذات الشروحات والتقييمات، وسنعود إليهما في القسم ١٠-٢-٤ أدناه.

أخيرًا، ومع تزايد استهلاك المستخدمين لمحتوى وسائل التواصل الاجتماعي على أجهزة مختلفة (كالحواسيب السطحية والأجهزة اللوحية والهواتف الذكية)، تبرز هناك حاجة لتطوير أساليب تتيح الوصول إلى المعلومات وتكون متوافقة مع منصات متعددة و/أو تكون مستقلة عن المنصات. لكن تصبح هذه المهمة صعبة بصفة خاصة عند عرض المعلومات في صيغة مرئية على الأجهزة ذات الشاشات الصغيرة.

١٠-٢-٢ الدمج والمعرفة الخلفية

تقليديًّا، تركز جهود الأبحاث على تطوير مسار بحثي معين، مثل الأساليب القائمة على القواعد أو أساليب التعلم الخاضعة للإشراف. تختلف مزايا المسارات البحثية، فبعضها يتميز في تعلم تمثيلات ونماذج الخصائص بناءً على بيانات تدريبية مصنفة، وتقديم التوقعات عن البيانات غير المرئية [60]، في حين يستفيد بعضها الآخر عن المعرفة الخلفية، على سبيل المثال، عن طريق تعلم قواعد الاستنباط بالاستناد إلى قواعد المعرفة الأولية (seed knowledge bases) [95، 110] أو إنشاء البيانات التدريبية تلقائيًّا لأغراض التعلم الخاضع للإشراف بالاعتماد على قواعد المعرفة الأولية (seed knowledge bases) [73، 81، 373].

من الأمور التي أثبتت فائدتها في العالم الحقيقي الحصول على وجهات نظر مختلفة عن المشكلة نفسها باستخدام أساليب مختلفة [95] أو باستخدام مخططات استخراج مختلفة [107] ودمجها معًا. ومع وجود بعض الأعمال التي أجريت في مجال دمج الأساليب المختلفة، على سبيل المثال استخدام تعلم المجموعات (learning ensemble) [374] أو المخططات الشاملة [107، 110]، مع الأخذ بالاعتبار أن غالبية الأعمال لا تركز على هذا الأمر. بالإضافة إلى ذلك، تفترض الأعمال التي تجرى في مجال دمج المخططات

وجود مخططين متداخلين. لكن في واقع الأمر، يُستخدم أكثر من مخططين اثنين لتعريف المعلومات. كما أن المخططات لا تكون متداخلة في جميع الأوقات، وهذا من الأسباب التي تدعو إلى استخدام مخططات مختلفة من البداية.

هناك تحديات إضافية لا تزال قائمة تتعلق بتعلم قواعد الاستنباط من قواعد المعرفة. في كثير من الأحيان، تأخذ أبحاث تعلم اللغات الطبيعية في الاعتبار الإعدادات الاصطناعية التي لا تُحدِّد فيها المخططات العلاقات القائمة بين المفاهيم أو الخصائص. على سبيل المثال، في نظام RDFS، تُحدَّد العلاقات بواسطة الخصائص التي توجد فيها خصائص فرعية ومجالات ونطاقات، بينما يسمح OWL بتعريف علاقات عكسية متبادلة. غير أن الأعمال التي تُعنى بتعلم الاستنباط تتجاهل ذلك إلى حد بعيد، وتفترض أنه ينبغي تعلم جميع العلاقات من هذا النوع بدءًا من الصفر، ولذا لا تركز على التحدي المتمثل في تجاوز نطاق ما جرى تعريفه مسبقًا.

١٠-٢-٣ قابلية التوسيع والفعالية

عندما يتعلق الأمر بأبحاث استخراج المعلومات، تعطي الخوارزميات ذات النطاق الكبير (يُشار إليها أيضًا باسم معالجة اللغات الطبيعية ذات البيانات الكثيفة أو على نطاق الويب) نتائج متفوقة مقارنة بنتائج المنهجيات التي تُدرّب على مجموعات بيانات أقل حجمًا [375]. يعود الفضل في ذلك إلى حد كبير إلى معالجة مشكلة تناثر البيانات عبر جمع أعداد أكبر بكثير من الأمثلة اللغوية التي تحدث بشكل طبيعي [375]. تشبه الحاجة إلى أساليب تعتمد على البيانات لإجراء عمليات معالجة اللغات الطبيعية ونجاح هذه الأساليب إلى حد بعيد الاتجاهات التي برزت في الآونة الأخيرة في المجالات البحثية الأخرى، وهذا يؤدي إلى ما يُشار إليه بعبارة «النموذج الرابع للعلم» (the fourth of science paradigm) [376].

في الوقت ذاته، ينبغي أن تكون عملية إضافة الشروحات الدلالية وخوارزميات الوصول إلى بيانات قابلة للتوسيع وفعالة، وذلك لكي تتكيف مع كميات البيانات الضخمة التي توجد في تدفقات وسائل التواصل الاجتماعي. تتطلب العديد من حالات الاستخدام معالجة إلكترونية شبه لحظية، وهو ما يبرز متطلبات إضافية من

حيث درجة تعقيد الخوارزمية. باتت الحوسبة السحابية [377] تعدُّ على نحو متزايد من العوامل الممكنة الأساسية التي تتيح إجراء عملية المعالجة بصورة قابلة للتوسيع وحسب الطلب، وهو ما يمنح الباحثين في أي مكان القدرة على الوصول إلى البنية التحتية الحوسبية بتكاليف ميسرة، ويسمح بتوفير طاقة حسابية كبيرة حسب الطلب ومن دون تكبد تكاليف مسبقة.

غير أن تطوير خوارزميات متوازية وقابلة للتوسيع لمنصات من قبيل Hadoop ليست مهمة سهلة على الإطلاق، لأن التشغيل والتبادل البسيط لمنظومات إضافة الشروحات الدلالية وموازاة الخوارزميات ليس سوى عدد قليل من المتطلبات التي ينبغي تلبيتها. ما زالت الأبحاث في هذا المجال في مراحلها الأولى، ولا سيّما تلك الأبحاث المتركزة حول منصات الأغراض العامة التي تختص بالمعالجة الدلالية القابلة للتوسيع.

يمكن اعتبار سحابة GATE [1] أنها الخطوة الأولى في هذا الاتجاه [320]. هذه المنصة الجديدة قائمة على الحوسبة السحابية لأبحاث تعدين النصوص واسعة النطاق، كما تدعم منظومات إضافة الشروحات الدلالية المبنية على الأنطولوجيات. تهدف هذه السحابة إلى تزويد الباحثين بمنصة كخدمة (platform-as-a-service)، وهو ما يتيح لهم إجراء اختبارات واسعة النطاق في مجال معالجة اللغات الطبيعية عبر استغلال الطاقة الحسابية الهائلة المتوفرة حسب الطلب على سحابة أمازون. كما تقلل الحاجة لتنفيذ خوارزميات مخصصة قابلة للموازاة. تتولى المنصة التعامل مع المشكلات البنيوية، وذلك بشكل شفاف تمامًا بالنسبة للباحث: موازنة الحمل، وتحميل البيانات وتخزينها بكفاءة، والتشغيل على الآلات الافتراضية، والأمان، وتدارك الأخطاء.

من الأمثلة على تطبيقات سحابة GATE أحد مشاريع الأرشيف الوطني البريطاني [293]، إذ جرى استخدامها لإضافة شروحات دلالية إلى ٤٢ تيرابايت من صفحات الويب وغيرها من المحتوى النصي. كانت عملية إضافة الشروحات مدعومة بواسطة قاعدة معرفية واسعة النطاق، مأخوذة من سحابة LOD، وموقع data.gov.uk،

1- http://cloud.gate.ac.uk

وقاعدة بيانات جغرافية ضخمة. جرت فهرسة النتائج في منصة GATE Mímir [297]، بالإضافة إلى واجهة مستخدم مخصصة للتصفح والبحث والانتقال من مساحة الوثيقة إلى قاعدة المعرفة الدلالية عبر بحث النص الكامل والشروحات الدلالية واستعلامات لغة «سباركل» (SPARQL).

١٠-٢-٤ التقييم ومجموعات البيانات المشتركة والتعهيد الجماعي

يعدُّ التقييم القضية المفتوحة الرابعة. وكما نوقش من قبل، قد يعيق انعدام معيار ذهبي مشترك لمجموعات البيانات إلى حد كبير قابلية التكرار والتقييم المقارن للخوارزميات. في الوقت ذاته، من المطلوب توفر تجارب تقييم معتمدة على المستخدمين أو مبنية على المهام، وذلك من أجل تحديد المشكلات الموجودة في أساليب البحث والعرض المرئي القائمة حاليًّا. هناك مجموعة كبيرة من الأبحاث التي لا تعرض نتائج اختبارات التقييم، أو الأبحاث التي قامت فقط بإجراء دراسات تكوينية ذات نطاق محدود، ولا سيّما في مجال الوصول المبتكر للمعلومات. على وجه الخصوص، هناك انعدام في عمليات التقييم الطولي (longitudinal evaluation) التي تجري بواسطة مجموعات مستخدمين أكبر حجمًا.

وبالمثل، يعد تدريب الخوارزميات وتكييفها على مجموعات البيانات التي تشكل المعيار الذهبي في وسائل التواصل الاجتماعي في الوقت الحالي محدودًا جدًّا. على سبيل المثال، لا توجد مجموعات بيانات المعيار الذهبي لتويتر وملخصات المدونات، كما يوجد أقل من ١٠,٠٠٠ تغريدة أضيفت إليها شروحات في صيغة كيانات أسماء. تعدُّ عملية إنشاء مجموعات بيانات كبيرة بما فيه الكفاية ولها ضرورة مهمة من خلال المنهجيات التقليدية المستندة إلى الخبراء لإضافة الشروحات النصية عملية باهظة الثمن، سواءٌ أكانت من حيث الوقت أم التمويل المطلوب، فقد يتراوح التمويل بين ٠,٣٦ دولار أمريكي و٠,١ دولار أمريكي للكلمة الواحدة [371]، وهو ما يعدُّ باهظ الثمن بالنسبة للمكانز المكونة من ملايين الكلمات. يمكن خفض التكاليف جزئيًّا عبر أدوات تعاونية متوفرة على الإنترنت لإضافة الشروحات، مثل أداة GATE Teamware [378] وأداة WebAnno [379]، وهما تدعمان أطقم عمل موزعة، كما أنهما تناسبان بشكل خاص مضيفي الشروحات غير الخبراء.

هناك بديل يشمل استخدام أسواق التعهيد الجماعي التجارية، إذ تشير التقارير إلى أن تكلفتها أقل بنسبة ٣٣٪ من تكلفة الموظفين التابعين للشركة، عندما يتعلق الأمر بإتمام مهام من قبيل تصنيف أقسام الكلام والتصنيف بشكل عام (classification) [380]. من ثم بدأ الباحثون في مجال معالجة اللغات إنشاء مكانز تحتوي على شروحات بواسطة خدمة «أمازون ميكانيكال تورك» (Amazon Mechanical Turk) وخدمة Crowdflower ومنهجيات أخرى معتمدة على الألعاب للحصول على وسائل بديلة أقل كلفة.

وبخصوص إضافة الشروحات إلى المكانز على وجه التحديد، تقدر دراسة (بويسيو وآخرون) [371] أنه مقارنة بتكلفة عمليات إضافة الشروحات التي تنفذ من قبل الخبراء (تقدر قيمتها بنحو مليون دولار)، يمكن تقليل تكلفة مليون من الوحدات اللغوية المضاف إليها الشروحات لما دون ٥٠٪ عبر استعمال خدمة «ميكانيكال تورك» (MTurk) (٣٨٠,٠٠٠ دولار - ٤٣٠,٠٠٠ دولار) لنحو ٢٠٪ (٢١٧,٩٢٧ دولار) عند استخدام منهجية مستندة إلى الألعاب مثل لعبة PhraseDetectives الخاصة بهذه الدراسة. وفيما يتعلق بإنشاء شروحات وسائل التواصل الاجتماعي عبر التعهيد الجماعي، كانت هناك بعض التجارب التي أجريت على تصنيف التغريدات إلى فئات [381] وإضافة الشروحات إلى كيانات الأسماء في التغريدات [292]، من بين أشياء أخرى. في مجال الويب الدلالي نفسه، استكشف الباحثون التعهيد الجماعي في الغالب عبر ألعاب هادفة، لاكتساب المعرفة في المقام الأول [382، 383] وتحسين LOD [384].

في الوقت ذاته، لجأ الباحثون إلى التعهيد الجماعي كوسيلة لتوسيع نطاق تجارب الاختبارات المستندة إلى العامل البشري. يكمن التحدي الرئيس هنا في كيفية تعريف مهمة التقييم، لكي يتسنى الحصول عليها عبر التعهيد الجماعي من أشخاص ليسوا خبراء، مع توفير نتائج عالية الجودة [385]. هذه المهمة ليست سهلة على الإطلاق، وقد جادل الباحثون بأن مهام التقييم التي تنفذ عبر التعهيد الجماعي ينبغي أن تُصمم بصورة مختلفة عن التقييمات التي تتم على أيدي الخبراء [386]. على وجه الخصوص،

خلصت دراسة جيليك وليو [386] إلى أن تقييم أنظمة التلخيص الذي يُنفذ من قبل أشخاص ليسوا خبراء يعطي نتائج مشوشة بصورة كبرى، ولذا فإنها تتطلب مزيدًا من التكرار للوصول إلى الأهمية الإحصائية (statistical significance)، كما أن عمال خدمة «ميكانيكال تورك» (Mechanical Turk) لا يمكنهم إعداد تصنيفات درجات متوافقة مع تصنيفات الخبراء.

من التصميمات الناجحة للتقييم المستند إلى التعهيد الجماعي تصميم يستخدم سير عمل مكون من أربع مراحل ذات مهام منفصلة، حيث جرى استخدامه في استيعاب قراءة الترجمة الآلية [367]. استُخدم تصميم أبسط للمهام في دراسة [387] لتقييم ملخصات التغريدات، حيث طُلب من عاملي موقع «ميكانيكال تورك» أن يحددوا، وفقًا لمقياس مكون من خمس نقاط، كمية المعلومات المنتجة من قبل البشر الموجودة في الملخص الذي جرى إنتاجه بصورة آلية. هناك مثال آخر من أمثلة التقييم، وهو مثال حقق نتائج ناجحة في موقع «ميكانيكال تورك»، وهو التصنيف المزدوج [388]. في هذه الحالة تكون المهمة تحديد الجملة الأكثر غنى بالمعلومات في أحد تقييمات المنتجات. في هذه الحالة، طُلب من عاملي التعهيد الجماعي ذكر ما إذا كانت الجملة التي اختيرت من قبل النظام المعياري تحمل قدرًا أكبر من المعلومات من جملة اختيرت بواسطة أسلوب المؤلف. جرى تحديد ترتيب الجمل بصورة عشوائية، وكان من الممكن أيضًا الإشارة إلى أن أيًّا من هذه الجمل كانت ملخصًا جيدًا. على الرغم من كل هذه الأعمال، لا تزال هناك مشكلات في أدوات تحويل المكانز القابلة للاستعمال المتكرر وواجهات المستخدم الخاصة بمهام معالجة اللغات الطبيعية التي تنفذ عبر التعهيد الجماعي. يعالج ملحق Gate Crowdsourcing للتعهيد الجماعي ذي المصدر المفتوح [389] هذا التحدي عبر توفير دعم بنيوي لمواءمة الوثائق مع وحدات التعهيد الجماعي والعكس، وذلك بصورة تلقائية، بالإضافة إلى التوليد التلقائي لواجهات تعهيد جماعي قابلة للاستعمال المتكرر لغرض إجراء مهام التصنيف والاختيار في عملية معالجة اللغات الطبيعية. يشار إلى أن سير العمل بأكمله قد جرى اختباره على مهام متنوعة من مهام معالجة اللغات الطبيعية، بما فيها إضافة الشروحات إلى كيانات الأسماء، وإزالة الغموض عن الكلمات، وكيانات

الأسماء فيما يتصل بمعرفات الموارد المميزة (URIs) الخاصة بقاعدة بيانات DBpedia، وإضافة الشروحات إلى أصحاب الآراء والأهداف، وكذلك المشاعر.

ختامًا، برز التعهيد الجماعي في الآونة الأخيرة كأسلوب واعد لإنشاء مجموعات بيانات تقييمية مشتركة، بالإضافة إلى تنفيذ اختبارات تقييم تُنفذ على يد المستخدمين. يعدُّ تكييف هذه الجهود لتتناسب مع الخصائص المحددة لعملية إضافة الشروحات الدلالية وعرض المعلومات في صيغة مرئية، بالإضافة إلى استخدامها لإنشاء موارد واسعة النطاق وتقييمات طولية قابلة للتكرار، من المجالات الأساسية لإجراء الأبحاث المستقبلية.

مسرد المصطلحات العلمية

المصطلح	المفهوم/ الترجمة
Stemming	اشتقاق جذع الكلمة
Aboutness	ارتباط النص بموضوع
Affixes	السوابق واللواحق (في الكلمات)
Aggregated analysis	التحليل الكلي
Annotation	إضافة التعليقات والشروحات
Approaches	منهجيات
Automatic Term Recognition	تمييز المصطلحات الآلي
Bag of words	كيس الكلمات
barrier word approach	منهجية كلمة الحاجز
Bigram	تسلسل العناصر الثنائي
boundary words	الكلمات الحدودية
Chunking	تجزئة النص
classifiers	المصنفات
Clustering	التجميع
Coarse-grained	تقريبي -إجمالي
collective analysis	التحليل الجماعي
computational linguistics (CL)	اللسانيات الحاسوبية
contradiction detection (CD)	ومهمة كشف التناقض
Controlled Language	لغة مقيّدة
Co-occurrence	التوارد المشترك

المصطلح	المفهوم/ الترجمة
Corpus	مكنز
crowdsourcing	الاشتراك الجماعي عبر الويب لتحقيق هدف أو حل مشكلة
data sparsity	تبعثر البيانات
dependency	تبعية – اعتماد
disambiguation	إزالة الغموض
domain-independent	ذو نطاق حر
Fine-grained	دقيق –تفصيلي
finite-state machine	آلة الحالات المحدودة
finite-state transducers	محولات طاقة محدودة
function words	الكلمات الوظيفية
Gazetteer	معجم كيانات الأسماء
gold-standard	معيار ذهبي
HMMs	نماذج ماركوف المخفية
human supervision	خاضع للإشراف البشري
Infix	الزوائد في أواسط الكلمات
Knowledge-based	المعتمد على المعرفة
Labelled	مصنف –مسمى
language-independent	مستقل اللغة
lemmas	إزالة الزوائد – المدخل المعجمي

المصطلح	المفهوم/ الترجمة
lexical analysis	التحليل المعجمي
maximum entropy	التحول الأقصى
Memes	فكرة أو صورة سريعة الانتشار على الويب
Metaphor	الاستعارات
Morpheme	أصغر وحدة لغوية ذات معنى
Morphological analysis	التحليل الصرفي
Morphology	الصرف
Named Entity	كيانات الأسماء
named entity classification (NEC)	تصنيف كيانات الأسماء واختصارها
Named entity linking (NEL)	ربط كيانات الأسماء
Named Entity Recognition (NER)	التعرف على كيانات الأسماء واختصارها
named entity recognition and classification (NERC)	مهمة التعرف على كيانات الأسماء وتصنيفها
natural language engineering (NLE)	هندسة اللغات الطبيعية
natural language generation (NLG)	توليد اللغات الطبيعية
Natural Language Processing (NLP)	معالجة اللغات الطبيعية
natural language understanding (NLU)	فهم اللغات الطبيعية
n-gram	تسلسل عدد من العناصر
Noisy	مشوشة-تشويش

المصطلح	المفهوم/ الترجمة
Nominals	اسمي -اعتباري
Normalization	تحويل النص إلى الشكل القياسي
Normalizer	محول النص للشكل القياسي
Noun Phrase	العبارة الاسمية
Ontology Design Patterns	أنماط تصميم الأنطولوجيات
Ontology Guided Information Extraction	استخلاص المعلومات الموجه بواسطة علم الأنماط
ontology learning and population (OLP)	عملية تعلم الأنماط والتعبئة
Ontology population	تعبئة الأنطولوجيا
Ontology-Based Information Extraction (OBIE)	استخلاص المعلومات المستندة إلى علم الأنماط
Opinion mining	تعدين الآراء
parameters	وسيط
Parser	محلل نحوي
Part-of-Speech (POS) tagging	تصنيف أقسام الكلام
Perceptrons	البيرسيبترونز: مستقبلات الشبكات العصبونية الاصطناعية، أحد خوارزميات التعلم الخاضع للإشراف
Pointwise Mutual Information	المعلومات المتبادلة الممثّلة بالنقاط
Polarity detection	كشف قطبية الرأي

المصطلح	المفهوم/ الترجمة
Predictive analysis	التحليل التنبئي
Prefix	السوابق (في الكلمات)
Question answering systems	أنظمة الإجابات على الأسئلة
recognizing textual entailment - RTE	تمييز الالتزام النصي
regular expression	التعبيرات القياسية
Relation Extraction	استخراج العلاقات
rule-based	المعتمد على القواعد
Seed	بذرة
Segmenting	تقطيع
semantic annotation	الشرح التوضيحي الدلالي
semantic drift	المغزى الدلالي
Semi- supervised	شبه خاضع للإشراف
Sentiment Analysis	تحليل المشاعر
Shallow or light parsing	التحليل السطحي
Splitter	مقسِّم
Suffix	اللواحق (في الكلمات)
Supervised	الخاضع للإشراف
Support Vector Machines (SVM)	آلات دعم المتجه
tagger	مصنف
Term extraction	استخراج المصطلحات

المفهوم/ الترجمة	المصطلح
تنقيب النصوص	Text mining
حد	Threshold
وحدة لغوية	token
تقطيع الكلمات	Tokenization
المعتمد على التحول	transformation-based
شجرة المعلومات	treebank
تسلسل العناصر الثلاثي	tri-gram
وحدة لغوية فريدة	Type
تسلسل العناصر الأحادي	Unigram
غير الخاضع للإشراف	unsupervised
معرف الموارد الموحد	URI
محدد الموارد المُوحّد	URL
سهم الاتجاه	Vector
العبارة الفعلية	Verb Phrase
جامع بيانات الويب	web crawler
مواقع تعاونية	Wiki
القاموس الحر التعاوني	Wiktionary
تضمين الكلمات	Word embeddings

المراجع

[1] Roger C. Schank and Larry Tesler. A conceptual dependency parser for natural language. In Proc. of the Conference on Computational Linguistics, pages 1–3. Association for Computational Linguistics, 1969. DOI: 10.3115/990403.990405. 2

[2] Robert B. Lees and N. Chomsky. Syntactic structures. Language, 33(3 Part 1), pages 375– 408, 1957. DOI: 10.2307/411160. 2

[3] R. Gaizauskas and Y. Wilks. Information extraction: Beyond document retrieval. Journal of Documentation, 54(1), pages 70–105, 1998. DOI: 10.1108/eum0000000007162. 2

[4] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. Gate: An architecture for development of robust hlt applications. In Proc. of the 40th Annual Meeting on Association for Computational Linguistics, ACL'02, pages 168–175, Stroudsburg, PA, 2002. DOI: 10.3115/1073083.1073112. 11

[5] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In Proc. of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 55–60, 2014. DOI: 10.3115/v1/p14-5010. 11

[6] Steven Bird, Ewan Klein, and Edward Loper. Natural Language Processing with Python. O'Reilly Media, Inc., 2009. 11

[7] H. Cunningham, D. Maynard, and V. Tablan. JAPE: A Java Annotation Patterns Engine 2nd ed. Research Memorandum CS–00–10, Department of Computer Science, University of Sheffield, Sheffield, UK, 2000. 14

[8] Tibor Kiss and Jan Strunk. Unsupervised multilingual sentence boundary detection. Computational Linguistics, 32(4), pages 485–525, 2006. DOI: 10.1162/coli.2006.32.4.485. 15

[9] Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. Building a large annotated corpus of english: The penn treebank. Computational Linguistics, 19(2), pages 313–330, 1994. 15

[10] W. Nelson Francis and Henry Kucera. Brown corpus manual. Brown University, 1979. 15

[11] Stig Johansson. The tagged {LOB} corpus: User\'s manual, 1986. 15

[12] E. Brill. A simple rule-based part-of-speech tagger. In Proc. of the 3rd Conference on Applied Natural Language Processing, Trento, Italy, 1992. DOI: 10.3115/974499.974526. 16

[13] Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, NAACL'03, pages 173–180, 2003. DOI: 10.3115/1073445.1073478. 16

[14] Thorsten Brants. Tnt: A statistical part-of-speech tagger. In Proc of the 6th conference on Applied Natural Language Processing, ANLP'00, pages 224–231, 2000. DOI: 10.3115/974147.974178. 16

[15] M. Hepple. Independence and commitment: Assumptions for rapid training and execution of rule-based part-of-speech taggers. In Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, Hong Kong, 2000. DOI: 10.3115/1075218.1075254. 16

[16] G. A. Miller. WordNet: An on-line lexical database. International Journal of Lexicography, 3(4), pages 235–312, 1990. 17

[17] M. F. Porter. An algorithm for suffix stripping. Program, 14(3), pages 130–137, 1980. DOI: 10.1108/eb046814. 19

[18] Ted Briscoe, John Carroll, and Rebecca Watson. The second release of the rasp system. In Proc. of the COLING/ACL on Interactive Presentation Sessions, pages 77–80, 2006. DOI: 10.3115/1225403.1225423. 19, 20

[19] D. Klein and C. Manning. Accurate unlexicalized parsing. In Proc. of the 41st Meeting of the Association for Computational Linguistics, 2003. DOI: 10.3115/1075096.1075150. 19, 21

[20] Robert Gaizauskas, Mark Hepple, Horacio Saggion, Mark A. Greenwood, and Kevin Humphreys. SUPPLE: A practical

parser for natural language engineering applications. In Proc. of the 9th International Workshop on Parsing Technology, pages 200–201. Association for Computational Linguistics, 2005. DOI: 10.3115/1654494.1654521. 19

[21] Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In Proc. of the International Conference on New Methods in Language Processing, volume 12, pages 44–49. Citeseer, 1994. 22

[22] L. Ramshaw and M. Marcus. Text chunking using transformation-based learning. In Proc. of the 3rd ACL Workshop on Very Large Corpora, 1995. DOI: 10.1007/978-94-017-2390- 9_10. 23

[23] Collins Cobuild, Ed. English Grammar. Harper Collins, 1999. 23

[24] S. Azar. Understanding and Using English Grammar. Prentice Hall Regents, 1989. 23

[25] R. Grishman and B. Sundheim. Message understanding conference-6: A brief history. In Proc. of COLING. Association for Computational Linguistics, 1995. DOI: 10.3115/992628.992709. 25, 26, 38

[26] Asif Ekbal, Eva Sourjikova, Anette Frank, and Simone Paolo Ponzetto. Assessing the challenge of fine-grained named entity recognition and classification. In A. Kumaran and Haizhou Li, Eds., Proc. of the Named Entities Workshop, pages 93–101, Uppsala, Sweden, 2010. Association for Computational Linguistics. 25

[27] Christopher Walker, Stephanie Strassel, Julie Medero, and Kazuaki Maeda. ACE 2005 multilingual training corpus. Linguistic Data Consortium, Philadelphia, 2006. 26, 27

[28] Erik F. Tjong, Kim Sang, and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In Walter Daelemans and Miles Osborne, Eds., Proc. of NAACL-HLT, pages 142–147, 2003. 27, 32

[29] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. Ontonotes: the 90% solution. In Proc. of the Human Language Technology Conference of the NAACL, Companion

Volume: Short Papers, pages 57–60, New York City, 2006. Association for Computational Linguistics. 27

[30] Giuseppe Rizzo and Raphaël Troncy. NERD: A framework for evaluating named entity recognition tools in the Web of data. In ISWC 10th International Semantic Web Conference, Bonn, Germany, 2011. 27

[31] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In Proc. of the 13th Conference on Computational Natural Language Learning, pages 147–155. Association for Computational Linguistics, 2009. DOI: 10.3115/1596374.1596399. 28

[32] Lev Ratinov and Dan Roth. Design challenges and misconceptions in named entity recognition. In Proc. of the 13th Conference on Computational Natural Language Learning, pages 147–155. Association for Computational Linguistics, 2009. DOI: 10.3115/1596374.1596399. 28

[33] James Pustejovsky, José M. Castano, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. Timeml: Robust specification of event and temporal expressions in text. New Directions in Question Answering, 3, pages 28– 34, 2003. 29

[34] J. Strötgen and M. Gertz. HeidelTime: High quality rule-based extraction and normalization of temporal expressions. In Proc. of the 5th International Workshop on Semantic Evaluation, pages 321–324. Association for Computational Linguistics, 2010. 29

[35] James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. ISO-TimeML: An international standard for semantic annotation. In LREC, 2010. 29

[36] Angel X. Chang and Christopher D. Manning. Sutime: A library for recognizing and normalizing time expressions. In LREC, pages 3735–3740, 2012. 29

[37] K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, and H. Cunningham. Shallow Methods for Named Entity Coreference Resolution. In Chaînes de Références et Résolveurs D'anaphores, Workshop TALN 2002, Nancy, France, 2002. http://gate.ac.uk/sale/taln02/taln-ws-coref.pdf 29

[38] Karthik Raghunathan, Heeyoung Lee, Sudarshan Rangarajan, Nathanael Chambers, Mihai Surdeanu, Dan Jurafsky, and Christopher Manning. A multi-pass sieve for coreference resolution. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 492–501. Association for Computational Linguistics, 2010. 29

[39] Roman Prokofyev, Alberto Tonon, Michael Luggen, Loic Vouilloz, Djellel Eddine Difallah, and Philippe Cudré-Mauroux. Sanaphor: Ontology-based coreference resolution. In the Semantic Web-ISWC 2015, pages 458–473. Springer, 2015. DOI: 10.1007/978-3- 319-25007-6_27. 29

[40] P. Cimiano, S. Handschuh, and S. Staab. Towards the self-annotating web. In Proc. of the 13th International Conference on World Wide Web (WWW '04), 2004. DOI: 10.1145/988672.988735. 30

[41] P. Pantel and M. Pennacchiotti. Automatically harvesting and ontologizing semantic relations. In Proc. of the Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pages 171–195. IOS Press, 2008. 30

[42] P. Cimiano, M. Hartung, and E. Ratsch. Learning the appropriate generalization level for relations extracted from the Genia corpus. In Proc. of the 5th Language Resources and Evaluation Conference (LREC), 2006. 30

[43] A. Schutz and P. Buitelaar. Relext: A tool for relation extraction from text in ontology extension. the Semantic Web-ISWC, pages 593–606, 2005. DOI: 10.1007/11574620_43. 30

[44] V. Crescenzi, G. Mecca, P. Merialdo, et al. Roadrunner: Towards automatic data extraction from large web sites. In Proc. of the International Conference on Very Large Data Bases, pages 109–118. Citeseer, 2001. 30

[45] D. E. Appelt. The common pattern specification language. Technical report, SRI International, Artificial Intelligence Center, 1996. DOI: 10.3115/1119089.1119095. 31

[46] D. Freitag. Information extraction from html: Application of a general learning approach. Proc. of the 15th Conference on Artificial Intelligence AAAI-98, pages 517–523, 1998. 31

[47] M. Califf and R. Mooney. Relational learning of pattern-match rules for information extraction. Working Papers of the ACL-97 Workshop in Natural Language Learning, pages 9– 15, 1997. 31

[48] S. Soderland. Learning information extraction rules for semi-structured and free text. Machine Learning, 34(1), pages 233–272, 1999. DOI: 10.1023/A: 1007562322031. 31

[49] Dayne Freitag and Nicholas Kushmerick. Boosted wrapper induction. In 17th National Conference on Artificial Intelligence (AAAI-2000): 12th Innovative Applications of Artificial Intelligence Conference (IAAI-2000), pages 577–583, 2000. 31

[50] F. Ciravegna. .LP/2, an adaptive algorithm for information extraction from web-related texts. In Proc. of the IJCAI Workshop on Adaptive Text Extraction and Mining, Seattle, 2001. 31

[51] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In Proc. of the 17th International Conference on Machine Learning, pages 591–598. Citeseer, 2000. 32

[52] H. L. Chieu and H. T. Ng. Named entity recognition with a maximum entropy approach. In Walter Daelemans and Miles Osborne, Eds., Proc. of CoNLL-2003, pages 160–163. Edmonton, Canada, 2003. 32

[53] H. Isozaki and H. Kazawa. Efficient support vector classifiers for named entity recognition. In Proc. of the 19th International Conference on Computational Linguistics (COLING'02), pages 390–396, Taipei, Taiwan, 2002. DOI: 10.3115/1072228.1072282. 32

[54] J. Mayfield, P. McNamee, and C. Piatko. Named entity recognition using hundreds of thousands of features. In Proc. of CoNLL-2003, pages 184–187. Edmonton, Canada, 2003. DOI: 10.3115/1119176.1119205. 32

[55] Y. Li, K. Bontcheva, and H. Cunningham. Using uneven margins SVM and perceptron for information extraction. In Proc.

of 9th Conference on Computational Natural Language Learning (CoNLL-2005), 2005. DOI: 10.3115/1706543.1706556. 32

[56] X. Carreras, L. Màrquez, and L. Padró. Learning a perceptron-based named entity chunker via online recognition feedback. In Proc. of CoNLL-2003, pages 156–159. Edmonton, Canada, 2003. DOI: 10.3115/1119176.1119198. 32

[57] Yaoyong Li, Kalina Bontcheva, and Hamish Cunningham. Adapting SVM for data sparseness and imbalance: A case study on information extraction. Natural Language Engineering, 15(2), pages 241–271, 2009. DOI: 10.1017/s1351324908004968. 32

[58] Andrew McCallum and Wei Li. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In Walter Daelemans and Miles Osborne, Eds., Proc. of the 7th Conference on Natural Language Learning at HLT- NAACL 2003, pages 188–191, 2003. DOI: 10.3115/1119176. 32

[59] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, Eds., Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics, pages 363–370, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. DOI: 10.3115/1219840. 32

[60] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. Journal of Machine Learning Research (JMLR), 999888, pages 2493–2537, 2011. 32, 137

[61] C. M. Bishop. Pattern Recognition and Machine Learning. Springer, New York, 2006. 32

[62] Kevin P. Murphy. Machine Learning: A Probabilistic Perspective. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2012. 32

[63] Xiao Ling and Daniel S. Weld. Fine-grained entity recognition. In Proc. of AAAI, pages 94–100. AAAI Press, 2012. 34

[64] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by Gibbs sampling. In Kevin Knight, Hwee Tou Ng, and Kemal Oflazer, Eds., Proc. of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05), pages 363–370, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. DOI: 10.3115/1219840. 34

[65] Colin Cherry and Hongyu Guo. The unreasonable effectiveness of word representations for twitter named entity recognition. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, Eds., Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 735–745, Denver, Colorado, 2015. Association for Computational Linguistics. DOI: 10.3115/v1/n15-1. 34

[66] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In Yuji Matsumoto and Rada Mihalcea, Eds., Proc. of the ACL-HLT, pages 368– 378, Portland, Oregon, 2011. Association for Computational Linguistics. 34

[67] Leon Derczynski, Diana Maynard, Giuseppe Rizzo, Marieke van Erp, Genevieve Gorrell, Raphaël Troncy, and Kalina Bontcheva. Analysis of named entity recognition and linking for tweets. Information Processing and Management, 51, pages 32–49, 2015. DOI: 10.1016/j.ipm.2014.10.006. 34, 59, 96

[68] Leon Derczynski, Isabelle Augenstein, and Kalina Bontcheva. USFD: Twitter NER with drift compensation and linked data. In Proc. of the 1st Workshop on Noisy Usergenerated Text. Association for Computational Linguistics, 2015. to appear. DOI: 10.18653/v1/w15-4306. 34

[69] Isabelle Augenstein, Leon Derczynski, and Kalina Bontcheva. Generalisation in named entity recognition: A quantitative analysis. Computer Speech and Language, 2016. under review. 35

[70] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional LSTM-CRF models for sequence tagging. CoRR, abs/1508.01991, 2015. 35

[71] Timothy Baldwin, Marie-Catherine de Marneffe, Bo Han, Young-Bum Kim, Alan Ritter, and Wei Xu. Shared tasks of the workshop on noisy user-generated text: Twitter lexical normalization and named entity recognition. In Wei Xu, Bo Han, and Alan Ritter, Eds., Proc. of the Workshop on Noisy User-generated Text, pages 126–135, Beijing, China, 2015. Association for Computational Linguistics. DOI: 10.18653/v1/w15-43. 35

[72] Ikuya Yamada, Hideaki Takeda, and Yoshiyasu Takefuji. Enhancing named entity recognition in twitter messages using entity linking. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, Eds., Proc. of the Workshop on Noisy User-generated Text, pages 136–140, Beijing, China, 2015. Association for Computational Linguistics. 35

[73] Isabelle Augenstein, Andreas Vlachos, and Diana Maynard. Extracting relations between non-standard entities using distant supervision and imitation learning. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 747–757, Lisbon, Portugal, 2015. Association for Computational Linguistics. DOI: 10.18653/v1/d15-1086. 37, 40, 45, 137

[74] Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. Injecting logical background knowledge into embeddings for relation extraction. In Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2015. DOI: 10.3115/v1/n15-1118. 37

[75] Mihai Surdeanu and Heng Ji. Overview of the english slot filling track at the TAC2014 knowledge base population evaluation. In Proc. of the TAC-KBP 2014 Workshop, 2014. 38, 40, 45

[76] Oren Etzioni, Anthony Fader, and Janara Christensen. Open information extraction: the second generation. In International Joint Conference on Artificial Intelligence (IJCAI), 2011. 39

[77] Rohit J. Kate and Raymond Mooney. Joint entity and relation extraction using card- pyramid parsing. In Mirella Lapata and Anoop Sarkar, Eds., Proc. of the 14th Conference on Computational Natural Language Learning, pages 203–212, Uppsala, Sweden, 2010. Association for Computational Linguistics. 40

[78] Sameer Singh, Sebastian Riedel, Brian Martin, Jiaping Zheng, and Andrew McCallum. Joint inference of entities, relations, and coreference. In Proc. of AKBC, pages 1–6. ACM, 2013. DOI: 10.1145/2509558.2509559. 40

[79] Qi Li and Heng Ji. Incremental joint extraction of entity mentions and relations. In Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 402–412, Baltimore, Maryland, 2014. Association for Computational Linguistics. DOI: 10.3115/v1/p14-1038. 40

[80] Heng Ji and Ralph Grishman. Knowledge base population: Successful approaches and challenges. In Yuji Matsumoto and Rada Mihalcea, Eds., Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1148– 1158, Portland, Oregon, 2011. Association for Computational Linguistics. 40

[81] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In Keh-Yih Su, Jian Su, Janyce Wiebe, and Haizhou Li, Eds., Proc. of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pages 1003–1011, Suntec, Singapore, 2009. Association for Computational Linguistics. DOI: 10.3115/1687878. 40, 45, 47, 48, 137

[82] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In Proc. of the ACM SIGMOD International Conference on Management of Data, pages 1247–1250. ACM, 2008. DOI: 10.1145/1376616.1376746. 41

[83] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A large ontology from wikipedia and WordNet. Web Semantics: Science, Services and Agents on the World Wide Web, 6(3), pages 203–217, 2008. DOI: 10.1016/j.websem.2008.06.001. 41

[84] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. Communications of the ACM, 57(10), pages 78–85, 2014. DOI: 10.1145/2629489. 41

[85] Sergey Brin. Extracting patterns and relations from the world wide web. In Paolo Atzeni, Alberto Mendelzon, and Giansalvatore Mecca, Eds., the World Wide Web and Databases, pages 172–183. Springer, 1999. DOI: 10.1007/10704656. 42

[86] Eugene Agichtein and Luis Gravano. Snowball: Extracting relations from large plain-text collections. In Peter Nürnberg, David Hicks, and Richard Furuta, Eds., Proc. of the 5th ACM Conference on Digital Libraries, pages 85–94, 2000. DOI: 10.1145/336597. 42

[87] O. Etzioni, M. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates. Web-scale information extraction in KnowItAll. In Stuart Feldman, Mike Uretsky, Marc Najork, and Craig Wills, Eds., Proc. of the 13th International Conference on World Wide Web, Rio de Janeiro, Brazil, 2004. ACM. 43

[88] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In Maria Fox and David Poole, Eds., Proc. of the 24th AAAI Conference on Artificial Intelligence, Palo Alto, California, 2010. AAAI Press. 43, 44

[89] Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In Proc. of the 14th International Conference on Computational Linguistics, pages 539–545, 1992. DOI: 10.3115/992133.992154. 43

[90] Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Coupled semi-supervised learning for information extraction. In Proc. of the 3rd ACM International Conference on Web Search and Data Mining, WSDM'10, pages 101–110, New York, NY, 2010. ACM. DOI: 10.1145/1718487.1718501. 44

[91] Saulo D. S. Pedro and Estevam R. Hruschka Jr. Conversing learning: Active learning and active social interaction for human supervision in never-ending learning systems. In Rubén Fuentes-Fernández Juan Pavón, Néstor D. Duque-Méndez, Ed., Advances in Artificial Intelligence—IBERAMIA 2012, pages 231–240. Springer, 2012. DOI: 10.1007/978- 3-642-34654-5. 44

[92] Stephen Soderland. Learning Text Analysis Rules for Domain Specific Natural Language Processing. Ph.D. thesis, University of Massachusetts, Amherst, MA, 1997. 44

[93] Warren Shen, AnHai Doan, Jeffrey F. Naughton, and Raghu Ramakrishnan. Declarative information extraction using datalog with embedded extraction predicates. In Christoph Koch, Johannes Gehrke, Minos N. Garofalakis, Divesh Srivastava, Karl Aberer, Anand Deshpande, Daniela Florescu, Chee Yong Chan, Venkatesh Ganti, Carl-Christian Kanne, Wolfgang Klas, and Erich J. Neuhold, Eds., VLDB, pages 1033–1044. ACM, 2007.

[94] Frederick Reiss, Sriram Raghavan, Rajasekar Krishnamurthy, Huaiyu Zhu, and Shivakumar Vaithyanathan. An algebraic approach to rule-based information extraction. In Proc. of the 24th IEEE International Conference on Data Engineering, pages 933–942. IEEE, 2008. DOI: 10.1109/icde.2008.4497502. 44

[95] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In Proc. of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 601–610. ACM, 2014. DOI: 10.1145/2623330.2623623. 44, 50, 137

[96] Aron Culotta and Jeffrey Sorensen. Dependency tree kernels for relation extraction. In Proc. of the 42nd Meeting of the Association for Computational Linguistics (ACL'04), Main Volume, pages 423–429, Barcelona, Spain, 2004. DOI: 10.3115/1218955.1219009. 45

[97] Razvan Bunescu and Raymond Mooney. A shortest path dependency kernel for relation extraction. In Raymond Mooney, Chris Brew, Program Co-chair Lee-Feng Chien, Academia Sinica, and Program Co-chair Katrin Kirchhoff, University of Washington, Eds., Proc. of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, pages 724–731, Vancouver, British Columbia, Canada, 2005. Association for Computational Linguistics. 45

[98] Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. Connecting language and knowledge bases with embedding models for relation extraction. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, Eds., Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 1366– 1371, Seattle, Washington, 2013. Association for Computational Linguistics. 45

[99] Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. TextRunner: Open information extraction on the Web. In Bob Carpenter, Amanda Stent, and Jason D. Williams, Eds., Proc. of Human Language Technologies: the Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 25–26, Rochester, New York, 2007. Association for Computational Linguistics. DOI: 10.3115/1614164. 46

[100] Anthony Fader, Stephen Soderland, and Oren Etzioni. Identifying relations for open information extraction. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard, Eds., Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 1535–1545, Seattle, Washington, 2013. Association for Computational Linguistics. 46

[101] Mausam, Michael Schmitz, Stephen Soderland, Robert Bart, and Oren Etzioni. Open language learning for information extraction. In Jun'ichi Tsujii, James Henderson, and Marius Pasça, Eds., Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 523–534, Jeju Island, Korea, 2012. Association for Computational Linguistics. 46

[102] Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. Leveraging linguistic structure for open domain information extraction. In Chengqing Zong and Michael Strube, Eds., Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference

on Natural Language Processing (Volume 1: Long Papers), pages 344–354, Beijing, China, 2015. Association for Computational Linguistics. DOI: 10.3115/v1/p15-1. 47

[103] Mark Craven, Johan Kumlien, et al. Constructing biological knowledge bases by extracting information from text sources. In thomas Lengauer, Reinhard Schneider, Peer Bork, Douglas Brutlag, Janice Glasgow, Hans-Werner Mewes, and Ralf Zimmer, Eds., Proc. of the International Conference on Intelligent Systems for Molecular Biology, volume 1999, pages 77–86, Palo Alto, California, 1999. AAAI Press. 47

[104] Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. Relation extraction from the Web using distant supervision. In Krzysztof Janowicz, Stefan Schlobach, Patrick Lambrix, and Eero Hyvönen, Eds., EKAW, volume 8876 of Lecture Notes in Computer Science, pages 26–41, Heidelberg, Germany, 2014. Springer. 48

[105] Isabelle Augenstein, Diana Maynard, and Fabio Ciravegna. Distantly supervised web relation extraction for knowledge base population. Semantic Web Journal, 7, 2016. DOI: 10.3233/sw-150180. 48

[106] Benjamin Roth, Tassilo Barth, Michael Wiegand, and Dietrich Klakow. A survey of noise reduction methods for distant supervision. In Fabian Suchanek, Sebastian Riedel, Sameer Singh, and Partha Pratim Talukdar, Eds., Proc. of the Workshop on Automated Knowledge Base Construction, pages 73–78, New York, NY, 2013. ACM. DOI: 10.1145/2505515.2505806. 48

[107] Sebastian Riedel, Limin Yao, Andrew McCallum, and Benjamin M. Marlin. Relation extraction with matrix factorization and universal schemas. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, Eds., Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 74–84, Atlanta, Georgia, 2013. Association for Computational Linguistics. 48, 137

[108] Sebastian Krause, Hong Li, Hans Uszkoreit, and Feiyu Xu. Large-scale learn ing of relation-extraction rules with distant supervision from the Web. In Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Manfred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist, Eds., International Semantic Web Conference (1), volume 7649 of Lecture Notes in Computer Science, pages 263–278. Springer, 2012. DOI: 10.1007/978-3-642- 35173-0. 49

[109] Gabor Angeli, Julie Tibshirani, Jean Wu, and Christopher D. Manning. Combining distant and partial supervision for relation extraction. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, Eds., Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1556–1567, Doha, Qatar, 2014. Association for Computational Linguistics. DOI: 10.3115/v1/d14-1. 49, 50

[110] Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. Injecting logical background knowledge into embeddings for relation extraction. In Rada Mihalcea, Joyce Chai, and Anoop Sarkar, Eds., Proc. of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1119–1129, Denver, Colorado, 2015. Association for Computational Linguistics. 49, 137

[111] D. Rao, P. McNamee, and M. Dredze. Entity linking: Finding extracted entities in a knowledge base. In Multi-source, Multi-lingual Inf. Extraction and Summarization. Springer, 2013. DOI: 10.1007/978-3-642-28569-1_5. 53, 59

[112] Silviu Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 708–716, 2007. 53

[113] A. Burman, A. Jayapal, S. Kannan, M. Kavilikatta, A. Alhelbawy, L. Derczynski, and R. Gaizauskas. USFD at KBP 2011: Entity linking, slot filling and temporal bounding. In Proc. of the Text Analysis Conference (TAC'11), 2011.

[114] D. Milne and I. H. Witten. Learning to link with wikipedia. In Proc. of the 17th Conference on Information and Knowledge Management (CIKM), pages 509–518, 2008. DOI: 10.1145/1458082.1458150. 53, 57, 95

[115] P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: Shedding light on the web of documents. In Proc. of I-SEMANTICS, pages 1–8, 2011. DOI: 10.1145/2063518.2063519. 53, 55, 95, 119

[116] J. Hoffart, M. A. Yosef, I. Bordino, H. Furstenau, M. Pinkal, M. Spaniol, B. Taneva, S. thater, and G. Weikum. Robust disambiguation of named entities in text. In Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 782–792, 2011. 53, 54, 57

[117] W. Shen, J. Wang, P. Luo, and M. Wang. LINDEN: Linking named entities with knowledge base via semantic knowledge. In Proc. of the 21st Conference on World Wide Web, pages 449–458, 2012. DOI: 10.1145/2187836.2187898. 53, 57

[118] Z.C. Zheng, X.C. Si, F.T. Li, E.Y. Chang, and X.Y. Zhu. Entity disambiguation with freebase. In Proc. of the Conference on Web Intelligence (WI-IAT'13), 2013. DOI: 10.1109/wi-iat.2012.26. 53

[119] G. Rizzo, R. Troncy, S. Hellmann, and M. Bruemmer. NERD meets NIF: Lifting NLP extraction results to the linked data cloud. In 5th Workshop on Linked Data on the Web (LDoW), 2012. 53, 58

[120] Ricardo Usbeck, Axel-Cyrille Ngonga Ngomo, Sören Auer, Daniel Gerber, and Andreas Both. Agdistis—graph-based disambiguation of named entities using linked data. In International Semantic Web Conference. 2014. DOI: 10.1007/978-3-319-11964-9_29. 53, 57

[121] E. Meij, W. Weerkamp, and M. de Rijke. Adding semantics to microblog posts. In Proc. of the 5th International Conference on Web Search and Data Mining (WSDM), 2012. DOI: 10.1145/2124295.2124364. 54, 55, 95

[122] Marco Cornolti, Paolo Ferragina, and Massimiliano Ciaramita. A framework for bench- marking entity-annotation systems. In Proc. of the 22nd International Conference on World Wide Web, WWW'13, pages 249–260, 2013. DOI: 10.1145/2488388.2488411. 54, 57, 59

[123] H. Ji, R. Grishman, H. T. Dang, K. Griffitt, and J. Ellis. Overview of the tac knowledge base population track. In Proc. of the 3rd Text Analysis Conference, 2010. 54

[124] H. Ji and R. Grishman. Knowledge base population: Successful approaches and challenges. In Proc. of ACL'2011, pages 1148–1158, 2011. 54

[125] Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. Discovering emerging entities with ambiguous names. In Proc. of the 23rd International Conference on World Wide Web, WWW'14, pages 385–396, 2014. DOI: 10.1145/2566486.2568003. 54, 57

[126] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In Proc. EMNLP, 2011. 55

[127] M. Rowe, M. Stankovic, A. S. Dadzie, B. P. Nunes, and A. E. Cano. Making sense of microposts (#msm2013): Big things come in small packages. In Proc. of the WWW Conference—Workshops, 2013. 55

[128] Amparo Elizabeth Cano Basave, Giuseppe Rizzo, Andrea Varga, Matthew Rowe, Milan Stankovic, and Aba-Sah Dadzie. Making sense of microposts (#microposts2014) named entity extraction and linking challenge. In 4th Workshop on Making Sense of Microposts (#Microposts2014), 2014. 55

[129] Genevieve Gorrell, Johann Petrak, and Kalina Bontcheva. Using @Twitter conventions to improve #lod-based named entity disambiguation. In the Semantic Web. Latest Advances and New Domains, pages 171–186. Springer, 2015. DOI: 10.1007/978-3-319-18818- 8_11. 55, 60, 97

[130] M. Michelson and S. A. Macskassy. Discovering users' topics of interest on Twitter: A first look. In Proc. of the 4th Workshop on Analytics for Noisy Unstructured Text Data, AND'10, pages 73–80, 2010. DOI: 10.1145/1871840.1871852. 55, 124

[131] David Milne and Ian H. Witten. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In Proc. of AAAI, 2008. 56

[132] Paolo Ferragina and Ugo Scaiella. Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In Proc. of the 19th ACM International Conference on Information and Knowledge Management, CIKM'10, pages 1625–1628, New York, NY, 2010. DOI: 10.1145/1871437.1871689. 57

[133] H. Saif, Y. He, and H. Alani. Alleviating data sparsity for Twitter sentiment analysis. In Proc. of the #MSM2012 Workshop, CEUR, volume 838, 2012. 58, 101

[134] F. Abel, Q. Gao, G. J. Houben, and K. Tao. Semantic enrichment of Twitter posts for user profile construction on the social web. In ESWC (2), pages 375–389, 2011. DOI: 10.1007/978-3-642-21064-8_26. 58, 59, 89, 96, 102, 122, 123, 124

[135] M. Rowe and M. Stankovic. Aligning tweets with events: Automation via semantics. Semantic Web, 1, 2009. DOI: 10.3233/SW-2011-0042. 58, 100

[136] S. Carter, W. Weerkamp, and E. Tsagkias. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. Language Resources and Evaluation Journal, 2013. DOI: 10.1007/s10579-012-9195-y. 59, 89, 96, 97

[137] U. Lösch and D. Müller. Mapping microblog posts to encyclopedia articles. Lecture Notes in Informatics, 192(150), 2011. 59

[138] Sherzod Hakimov, Salih Atilay Oto, and Erdogan Dogdu. Named entity recognition and disambiguation using linked data and graph-based centrality scoring. In Proc. of the 4th International Workshop on Semantic Web Information Management, 2012. DOI: 10.1145/2237867.2237871. 59

[139] Wei Shen, Jianyong Wang, Ping Luo, and Min Wang. Linking named entities in tweets with knowledge base via user interest modeling. In Proc. of the 19th ACM SIGKDD International Conference

on Knowledge Discovery and Data Mining, pages 68–76. ACM, 2013. DOI: 10.1145/2487575.2487686. 59, 96

[140] Hongzhao Huang, Yunbo Cao, Xiaojiang Huang, Heng Ji, and Chin-Yew Lin. Collective tweet wikification based on semi-supervised graph regularization. In Proc. of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 380–390, 2014. DOI: 10.3115/v1/p14-1036. 59, 96

[141] Abhishek Gattani, Digvijay S Lamba, Nikesh Garera, Mitul Tiwari, Xiaoyong Chai, San- jib Das, Sri Subramaniam, Anand Rajaraman, Venky Harinarayan, and AnHai Doan. Entity extraction, linking, classification, and tagging for social media: A wikipedia-based approach. Proc. of the VLDB Endowment, 6(11), pages 1126–1137, 2013. DOI: 10.14778/2536222.2536237. 59, 96

[142] Jens Lehmann and Johanna Völker. Perspectives on Ontology Learning, volume 18. IOS Press, 2014. 61

[143] Paul Buitelaar and Philipp Cimiano. Ontology Learning and Population: Bridging the Gap Between Text and Knowledge, volume 167. IOS Press, 2008.

[144] P. Buitelaar, P. Cimiano, and B. Magnini. Ontology Learning from Text: Methods, Applications and Evaluation. IOS Press, 2005. 61

[145] T. Berners-Lee, D. Connolly, and R. R. Swick. Web architecture: Describing and exchanging data. Technical report, W3C Consortium,

http://www.w3.org/\protect\discretionary{\char\hyphenchar\font}{}{}1999/04/WebData, 1999. 62

[146] Nitin Indurkhya and Fred J. Damerau. Handbook of Natural Language Processing, volume 2. CRC Press, 2010. 63

[147] G. Salton and M. J. McGill. Introduction to Modern Information Retrieval. McGraw-Hill, 1983. 64

[148] W. Bosma and P. Vossen. Bootstrapping language-neutral term extraction. In 7th Language Resources and Evaluation Conference (LREC), Valletta, Malta, 2010. 65

[149] K. T. Frantzi and S. Ananiadou. the C-Value/NC-Value domain independent method for multi-word term extraction. Journal of Natural Language Processing, 6(3), pages 145–179, 1999. DOI: 10.5715/jnlp.6.3_145. 65

[150] D. G. Maynard and S. Ananiadou. Identifying terms by their family and friends. In Proc. of 18th International Conference on Computational Linguistics (COLING), Saarbrücken, Germany, 2000. DOI: 10.3115/990820.990897. 65

[151] D. Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In Proc. of 14th International Conference on Computational Linguistics (COLING), pages 977–981, Nantes, France, 1992. DOI: 10.3115/992383.992415. 65

[152] S. J. Nelson, N. E. Olson, L. Fuller, M. S. Tuttle, W. G. Cole, and D. D. Sherertz. Identifying concepts in medical knowledge. In Proc. of 8th World Congress on Medical Informatics (MEDINFO), pages 33–36, 1995. 65

[153] Alexander Maedche and Steffen Staab. Ontology learning. In Handbook on Ontologies, pages 173–190. Springer, 2004. DOI: 10.1007/978-3-540-24750-0_9. 66

[154] Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. Communications of the ACM, 18(11), pages 613–620, 1975. DOI: 10.1145/361219.361220. 66

[155] G. Heyer and H. F. Witschel. Terminology and metadata—on how to effciently build an ontology. TermNet News—Newsletter of International Cooperation in Terminology, 87, 2005. 66

[156] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Comparing conceptual, divise and agglomerative clustering for learning taxonomies from text. In Proc. of the 16th European Conference on Artificial Intelligence, ECAI'2004, Including Prestigious Applicants of Intelligent Systems, PAIS, 2004. 66

[157] Diana Maynard. Term Recognition Using Combined Knowledge Sources. Ph.D. thesis, Department of Computing and Mathematics, Manchester Metropolitan University, UK, 1999. 66, 67

[158] G. Grefenstette. Explorations in Automatic thesaurus Discovery. Kluwer Academic Publishers, 1994. DOI: 10.1007/978-1-4615-2710-7. 66

[159] G. Rigau, J. Atserias, and E. Agirre. Combining unsupervised lexical knowledge methods for word sense disambiguation. In Proc. of ACL/EACL, pages 48–55, Madrid, Spain, 1997. DOI: 10.3115/976909.979624. 67

[160] A. Smeaton and I. Quigley. Experiments on using semantic distances between words in image caption retrieval. In Proc. of 19th International Conference on Research and Development in Information Retrieval, Zurich, Switzerland, 1996. DOI: 10.1145/243199.243261. 67

[161]Gang Zhao. Analogical Translator: Experience-Guided Transfer in Machine Translation. Ph.D. thesis, Department of Language Engineering, UMIST, Manchester, England, 1996. 67

[162] T. Tsutsumi. Natural language processing: the PLNLP approach. In K. Jenon, G. E Heidhorn, and S. D. Richardson, Eds., Word Sense Disambiguation by Examples, pages 263–272. Kluwer Academic Publishers, Dordrecht, 1993. 67

[163] N. Uramoto. A best-match algorithm for broad-coverage example-based disambiguation. In Proc. of 15th International Conference on Computational Linguistics, volume 2, pages 717– 721, Kyoto, Japan, 1994. DOI: 10.3115/991250.991261. 67

[164] E. Sumita and H. Iida. Experiments and prospects of example-based machine translation. In Proc. of 29th Annual Meeting of the Association for Computational Linguistics, pages 185– 192, Berkeley, California, 1991. DOI: 10.3115/981344.981368. 67

[165] M. A. Hearst. Automatic acquisition of hyponyms from large text corpora. In Conference on Computational Linguistics (COLING'92), Nantes, France, 1992. Association for Computational Linguistics. DOI: 10.3115/992133.992154. 68

[166] M. Berland and E. Charniak. Finding parts in very large corpora. In Proc. of ACL-99, pages 57–64, College Park, MD, 1999. DOI: 10.3115/1034678.1034697. 68

[167] Z. S. Harris. Mathematical Structures of Language. Wiley (Interscience), New York, 1968. 68, 69

[168] Frank Smadja. Retrieving collocations from text: Xtract. Computational Linguistics, 19(1), pages 143–177, 1993. 68

[169] Wim Peters. Text-based legal ontology enrichment. Proc. of LOAIT, pages 55–66, 2009. 68

[170] Naomi Sager. Syntactic formatting of scientific information. In Proc. of 1972 Fall Joint Computer Conference, volume 41 of AFIPS Conf. Proc., pages 791–800, Montvale, NJ, 1972. DOI: 10.1145/1480083.1480101. 69

[171] L. Hirschman, R. Grishman, and N. Sager. Grammatically based automatic word class formation. Information Processing and Retrieval, 11, pages 39–57, 1975. DOI: 10.1016/0306- 4573(75)90033-3. 69

[172] L. Hirschman and N. Sager. Automatic information formatting of a medical sublanguage. In Kittredge and Lehrberger, Eds., Sublanguage: Studies of Language in Restricted Semantic Domains, pages 27–69. Walter de Gruyter, 1982. DOI: 10.1515/9783110844818. 69

[173] R. A. Rocha, B. Rocha, and S. M. Huff. Automated translation between medical vocabularies using a frame-based interlingua. In Proc. of SCAMC'94, pages 690–694, 1994. 70

[174] P. Cimiano and J. Voelker. Text2Onto—A framework for ontology learning and data- driven change discovery. In Proc. of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB), Alicante, Spain, 2005. 70

[175] D. Maynard, A. Funk, and W. Peters. SPRAT: A tool for automatic semantic pattern- based ontology population. In International Conference for Digital Libraries and the Semantic Web, Trento, Italy, 2009. 70

[176] Francesco Draicchio, Aldo Gangemi, Valentina Presutti, and Andrea Giovanni Nuzzolese. FRED: From natural language text to

RDF and owl in one click. In Extended Semantic Web Conference, pages 263–267. Springer, 2013. DOI: 10.1007/978-3-642-41242-4_36. 70

[177] Johan Bos. Wide-coverage semantic analysis with boxer. In Proc. of the Conference on Semantics in Text Processing, pages 277–286. Association for Computational Linguistics, 2008. DOI: 10.3115/1626481.1626503. 70

[178] Aldo Gangemi. Ontology design patterns for semantic web content. In the Semantic Web-ISWC, pages 262–276. Springer, 2005. DOI: 10.1007/11574620_21. 71

[179] G. Aguade de Cea, A. Gómez-Pérez, E. Montiel Ponsoda, and M-C. Suárez-Figueroa. Natural language-based approach for helping in the reuse of ontology design patterns. In Proc. of the 16th International Conference on Knowledge Engineering and Knowledge Management Knowledge Patterns (EKAW), Acitrezza, Italy, 2008. DOI: 10.1007/978-3-540- 87696-0_6. 71

[180] Kaarel Kaljurand and Norbert E Fuchs. Verbalizing OWL in attempto controlled english. In OWLED, volume 258, 2007. DOI: 10.5167/uzh-33256. 71

[181] Cathy Dolbear, Glen Hart, Katalin Kovacs, John Goodwin, and Sheng Zhou. the rabbit language: Description, syntax and conversion to OWL. Ordinance Survey Research Labs Technical Report, 2007. 71

[182] Anne Cregan, Rolf Schwitter, thomas Meyer, et al. Sydney owl syntax-towards a con- trolled natural language syntax for owl 1.1. In OWLED, volume 258, 2007. 71

[183] Adam Funk, Valentin Tablan, Kalina Bontcheva, Hamish Cunningham, Brian Davis, and Siegfried Handschuh. CLOnE: Controlled language for ontology editing. In Proc. of the 6th International Semantic Web Conference (ISWC), Busan, Korea, 2007. DOI: 10.1007/978-3-540-76298-0_11. 71

[184] Diana Maynard and Mark A. Greenwood. Who cares about sarcastic tweets? Investigating the impact of sarcasm on sentiment analysis. In Proc. of LREC, Reykjavik, Iceland, 2014. 75, 78, 79

[185] Diana Maynard and Jonathon Hare. Entity-based opinion mining from text and multimedia. In Advances in Social Media Analysis, pages 65–86. Springer, 2015. DOI: 10.1007/978- 3-319-18458-6_4. 77

[186] Xiaowen Ding, Bing Liu, and Lei Zhang. Entity discovery and assignment for opinion mining applications. In Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 1125–1134. ACM, 2009. DOI: 10.1145/1557019.1557141. 77

[187] Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In Proc. of the International Workshop on Semantic Evaluation, SemEval'16, San Diego, California, 2016. DOI: 10.18653/v1/s16-1003. 77

[188] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding. In Proc. of EMLNP, 2016. 77

[189] Leonardo Rocha, Fernando Mourão, thiago Silveira, Rodrigo Chaves, Giovanni Sá, Felipe Teixeira, Ramon Vieira, and Renato Ferreira. Saci: Sentiment analysis by collective inspection on social media content. Web Semantics: Science, Services and Agents on the World Wide Web, 2015. DOI: 10.1016/j.websem.2015.05.006. 78

[190] Diana Maynard, Gerhard Gossen, Marco Fisichella, and Adam Funk. Should I care about your opinion? Detection of opinion interestingness and dynamics in social media. Journal of Future Internet, 2015. DOI: 10.3390/fi6030457. 78

[191] Christine Liebrecht, Florian Kunneman, and Antal van den Bosch. the perfect solution for detecting sarcasm in tweets# not. WASSA, page 29, 2013. 78

[192] David Bamman and Noah A Smith. Contextualized sarcasm detection on twitter. In 9th International AAAI Conference on Web and Social Media, 2015. 79

[193] Ameeta Agrawal and Aijun An. Unsupervised emotion detection from text using semantic and syntactic relations. In Proc. of the IEEE/WIC/ACM International Joint Conferences on Web Intelligence and

Intelligent Agent Technology, Volume 01, WI-IAT'12, pages 346–353, Washington, DC, 2012. IEEE Computer Society. DOI: 10.1109/wi-iat.2012.170. 79

[194] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. http://arxiv.org/abs/0911.1583, 2009. 79, 100

[195] Marc Schröder, Paolo Baggia, Felix Burkhardt, Catherine Pelachaud, Christian Peter, and Enrico Zovato. EmotionML—an upcoming standard for representing emotions and related states. In Affective Computing and Intelligent Interaction, pages 316–325. Springer, 2011. DOI: 10.1007/978-3-642-24600-5_35. 79

[196] W. Gerrod Parrott. Emotions in Social Psychology: Essential Readings. Psychology Press, 2001. 79

[197] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: further exploration of a prototype approach. Journal of Personality and Social Psychology, 52(6), page 1061, 1987. DOI: 10.1037//0022-3514.52.6.1061. 79

[198] B. Pang and L. Lee. Opinion mining and sentiment analysis. Information Retrieval, 2(1), 2008. DOI: 10.1561/1500000011. 81

[199] S. Moghaddam and F. Popowich. Opinion polarity identification through adjectives. CoRR, abs/1011.4623, 2010. 82

[200] A. C. Mullaly, C. L. Gagné, T. L. Spalding, and K. A. Marchak. Examining ambiguous adjectives in adjective-noun phrases: Evidence for representation as a shared core-meaning. the Mental Lexicon, 5(1), pages 87–114, 2010. DOI: 10.1075/ml.5.1.04mul. 82

[201] A. Weichselbraun, S. Gindl, and A. Scharl. A context-dependent supervised learning approach to sentiment detection in large textual databases. Journal of Information and Data Management, 1(3), pages 329–342, 2010. 83

[202] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational Linguistics, 35(3), pages 399– 433, 2009. DOI: 10.1162/coli.08-012-r1-06-90. 83

[203] S. Gindl, A. Weichselbraun, and A. Scharl. Cross-domain contextualisation of sentiment lexicons. In Proc. of 19th European Conference on Artificial Intelligence (ECAI), pages 771– 776, 2010. DOI: 10.3233/978-1-60750-606-5-771. 83

[204] A. Aue and M. Gamon. Customizing sentiment classifiers to new domains: A case study. In Proc. of the International Conference on Recent Advances in Natural Language Processing, Borovetz, Bulgaria, 2005. 83

[205] Krisztian Balog, Gilad Mishne, and Maarten De Rijke. Why are they excited? Identifying and explaining spikes in blog mood levels. In Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics: Posters and Demonstrations, pages 207–210. Association for Computational Linguistics, 2006. DOI: 10.3115/1608974.1609010. 83

[206] C. J. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In 8th International AAAI Conference on Weblogs and Social Media, 2014. 83

[207] M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. Lexicon-based methods for sentiment analysis. Computational Linguistics, 1(September 2010), pages 1–41, 2011. DOI: 10.1162/coli_a_00049. 83

[208] Diego Reforgiato Recupero, Mauro Dragoni, and Valentina Presutti. Eswc'15 challenge on concept-level sentiment analysis. In Semantic Web Evaluation Challenge, pages 211–222. Springer, 2015. DOI: 10.1007/978-3-319-12024-9_1. 84

[209] Diego Reforgiato Recupero and Erik Cambria. Eswc'14 challenge on concept-level sentiment analysis. In Semantic Web Evaluation Challenge, pages 3–20. Springer, 2015. DOI: 10.1007/978-3-319-12024-9_1. 84

[210] Anni Coden, Dan Gruhl, Neal Lewis, Pablo N. Mendes, Meena Nagarajan, Cartic Ramakrishnan, and Steve Welch. Semantic lexicon expansion for concept-based aspect-aware sentiment analysis. In Semantic Web Evaluation Challenge, pages 34–40. Springer, 2014. DOI: 10.1007/978-3-319-12024-9_4. 84

[211] Pablo N. Mendes, Max Jakob, and Christian Bizer. Dbpedia: A multilingual cross-domain knowledge base. In LREC, pages 1813–1817, 2012. 84

[212] Pei Yin, Hongwei Wang, and Kaiqiang Guo. Feature—opinion pair identification of product reviews in chinese: A domain ontology modeling method. New Review of Hypermedia and Multimedia, 19(1), pages 3–24, 2013. DOI: 10.1080/13614568.2013.766266. 84

[213] Samaneh Moghaddam and Martin Ester. the flda model for aspect-based opinion mining: addressing the cold start problem. In Proc. of the 22nd international conference on World Wide Web, pages 909–918. International World Wide Web Conferences Steering Committee, 2013. DOI: 10.1145/2488388.2488467. 84

[214] Mike thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), pages 2544–2558, 2010. DOI: 10.1002/asi.21416. 85, 119

[215] Peter Pirolli. Powers of 10: Modeling complex information-seeking systems at multiple scales. IEEE Computer, 42(3), pages 33–40, 2009. DOI: 10.1109/mc.2009.94. 87

[216] N. Ravikant and A. Rifkin. Why twitter is massively undervalued compared to facebook. TechCrunch, 2010. http://techcrunch.com/2010/10/16/why-twitter-is- massively-undervalued-compared-to-facebook/ 88

[217] Meredith M. Skeels and Jonathan Grudin. When social networks cross boundaries: A case study of workplace use of facebook and linkedIn. In Proc. of the ACM International Conference on Supporting Group Work, GROUP'09, pages 95–104, New York, NY, 2009. DOI: 10.1145/1531674.1531689. 88

[218] K. Bontcheva and H. Cunningham. Semantic annotation and retrieval: Manual, semiautomatic and automatic generation. In J. Domingue, D. Fensel, and J. A. Hendler, Eds., Handbook of Semantic Web Technologies. Springer, 2011. DOI: 10.1007/978-3-540-92913- 0. 88

[219] X. Liu, S. Zhang, F. Wei, and M. Zhou. Recognizing named entities in tweets. In Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 359–367, 2011. 88, 96

[220] A. Ritter, S. Clark, Mausam, and O. Etzioni. Named entity recognition in tweets: An experimental study. In Proc. of Empirical Methods for Natural Language Processing (EMNLP), Edinburgh, UK, 2011. 88, 93, 95, 97, 98

[221] P. N. Mendes, A. Passant, P. Kapanipathi, and A. P. Sheth. Linked open social signals. In Proc. of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT'10, pages 224–231, Washington, DC, 2010. IEEE Computer Society. DOI: 10.1109/wi-iat.2010.314. 89, 95, 119

[222] B. Han and T. Baldwin. Lexical normalisation of short text messages: Makn sens a #twitter. In Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT'11, pages 368–378, 2011. 89

[223] S. Gouws, D. Metzler, C. Cai, and E. Hovy. Contextual bearing on linguistic variation in social media. In Proc. of the Workshop on Languages in Social Media, LSM'11, pages 20–29, 2011. 89

[224] T. Baldwin and M. Lui. Language identification: the long and the short of the matter. In Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 229–237, Los Angeles, California, 2010. 89

[225] I. Celino, D. Dell'Aglio, E. Della Valle, Y. Huang, T. Lee, S. Park, and V. Tresp. Making sense of location-based micro-posts using stream reasoning. In Proc. of the Making Sense of Microposts Workshop (#MSM2011), Collocated with the 8th Extended Semantic Web Conference, Heraklion, Crete, Greece, 2011. 91

[226] S. Scerri, K. Cortis, I. Rivera, and S. Handschuh. Knowledge Discovery in Distributed Social Web Sharing Activities. In Proc. of the #MSM2012 Workshop, CEUR, volume 838, 2012. 91

[227] T. Plumbaum, S. Wu, E. W. De Luca, and S. Albayrak. User Modeling for the Social Semantic Web. In 2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, in conjunction with ISWC, 2011. 91

[228] A. Passant and P. Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In Proc. of the Linked Data on the Web Workshop (LDOW), Beijing, China, 2008. 91

[229] A. Passant, J. G. Breslin, and S. Decker. Rethinking microblogging: Open, distributed, semantic. In Proc. of the 10th International Conference on Web Engineering, pages 263–277, 2010. DOI: 10.1007/978-3-642-13911-6_18. 91

[230] Dominik Heckmann, Tim Schwartz, Boris Brandherm, Michael Schmitz, and Margeritta von Wilamowitz-Moellendorff. GUMO—the general user model ontology. In Proc. of the 10th International Conference on User Modeling, pages 428–432, 2005. DOI: 10.1007/11527886_58. 91

[231] S. Angeletou, M. Rowe, and H. Alani. Modelling and analysis of user behaviour in online communities. In Proc. of the 10th International Conference on the Semantic Web, ISWC'11, pages 35–50. Springer-Verlag, 2011. DOI: 10.1007/978-3-642-25073-6_3. 92, 123, 125

[232] M. Rowe, S. Angeletou, and H. Alani. Predicting discussions on the social semantic web. In Proc. of the 8th Extended Semantic Web Conference on the Semanic Web, ESWC'11, pages 405–420. Springer-Verlag, 2011. DOI: 10.1007/978-3-642-21064-8_28. 92, 123

[233] R. Mihalcea and P. Tarau. TextRank: Bringing order into text. In Proc. of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 404–411, 2004. 93

[234] W. Wu, B. Zhang, and M. Ostendorf. Automatic generation of personalized annotation tags for twitter users. In Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 689–692, 2010. 93, 125

[235] F. Zanzotto, M. Pennaccchiotti, and K. Tsioutsiouliklis. Linguistic Redundancy in Twitter. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 659– 669, Edinburgh, UK, 2011. Association for Computational Linguistics. 93

[236] B. Sharifi, M. A. Hutton, and J. Kalita. Summarizing microblogs automatically. In Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 685–688, Los Angeles, California, 2010. 93

[237] W. Xin, Z. Jing, J. Jing, H. Yang, S. Palakorn, W. X. Zhao, J. Jiang, J. He, Y. Song, P. Achananuparp, E. P. Lim, and X. Li. Topical keyphrase extraction from twitter. In Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT'11, pages 379–388, 2011. 93, 125

[238] Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In Proc. of the 4th International Conference on Weblogs and Social Media (ICWSM), 2010. 93, 132

[239] G. Mishne. AutoTag: A collaborative approach to automated tag assignment for weblog posts. In Proc. of the 15th International Conference on World Wide Web, pages 953–954, 2006. DOI: 10.1145/1135777.1135961. 93

[240] L. Qu, C. Müller, and I. Gurevych. Using tag semantic network for keyphrase extraction in blogs. In Proc. of the 17th Conference on Information and Knowledge Management, pages 1381–1382, 2008. DOI: 10.1145/1458082.1458290. 93

[241] G. Solskinnsbakk and J. A. Gulla. Semantic annotation from social data. In Proc. of the 4th International Workshop on Social Data on the Web Workshop, 2011. 93

[242] N. Ireson and F. Ciravegna. Toponym resolution in social media. In Proc. of the 9th International Semantic Web Conference (ISWC), pages 370–385, 2010. DOI: 10.1007/978- 3-642-17746-0_24. 95

[243] David Laniado and Peter Mika. Making sense of twitter. In Peter Patel-Schneider, Yue Pan, Pascal Hitzler, Peter Mika, Lei Zhang, Jeff

Pan, Ian Horrocks, and Birte Glimm, Eds., the Semantic Web (ISWC), volume 6496 of Lecture Notes in Computer Science, pages 470–485. Springer Berlin/Heidelberg, 2010. 95

[244] D. Gruhl, M. Nagarajan, J. Pieper, C. Robson, and A. Sheth. Context and domain knowledge enhanced entity spotting in informal text. In Proc. of the 8th International Semantic Web Conference (ISWC), 2009. DOI: 10.1007/978-3-642-04930-9_17. 95, 104

[245] S. Choudhury and J. Breslin. Extracting semantic entities and events from sports tweets. In Proc. of the 1st Workshop on Making Sense of Microposts (MSM): Big things Come in Small Packages, pages 22–32, 2011. 100

[246] Elizabeth L. Murnane, Bernhard Haslhofer, and Carl Lagoze. Reslve: Leveraging user interest to improve entity disambiguation on short text. In Proc. of the 22nd International Conference on World Wide Web, pages 1275–1284, 2013. DOI: 10.1145/2487788.2488162. 96

[247] Chenliang Li, Jianshu Weng, Qi He, Yuxia Yao, Anwitaman Datta, Aixin Sun, and BuSung Lee. Twiner: Named entity recognition in targeted twitter stream. In Proc. of the 35th ACM Conference on Research and Development in Information Retrieval, pages 721– 730. ACM, 2012. DOI: 10.1145/2348283.2348380. 96

[248] Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A. Greenwood, Diana Maynard, and Niraj Aswani. TwitIE: An open-source information extraction pipeline for microblog text. In Proc. of the International Conference on Recent Advances in Natural Language Processing. Association for Computational Linguistics, 2013. 97, 134

[249] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, N. Aswani, I. Roberts, G. Gorrell, A. Funk, A. Roberts, D. Damljanovic, T. Heitz, M. A. Greenwood, H. Saggion, J. Petrak, Y. Li, and W. Peters. Text Processing with GATE (Version 6). 2011. 97

[250] B. Han, P. Cook, and T. Baldwin. Automatically constructing a normalisation dictionary for microblogs. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 421–432. ACL, 2012. 98

[251] L. Derczynski, D. Maynard, N. Aswani, and K. Bontcheva. Microblog-genre noise and impact on semantic annotation accuracy. In Proc. of the 24th ACM Conference on Hypertext and Social Media, 2013. DOI: 10.1145/2481492.2481495. 98

[252] E. Forsyth and C. Martell. Lexical and discourse analysis of online chat dialog. In International Conference on Semantic Computing, pages 19–26. IEEE, 2007. DOI: 10.1109/icosc.2007.4338328. 99

[253] M. P. Marcus, B. Santorini, and M. A. Marcinkiewicz. Building a Large Annotated Corpus of English: the Penn Treebank. Computational Linguistics, 19(2), pages 313–330, 1993. 99

[254] M. Dork, D. Gruen, C. Williamson, and S. Carpendale. A visual backchannel for large-scale events. IEEE Transactions on Visualization and Computer Graphics, 16(6), pages 1129–1138, 2010. DOI: 10.1109/tvcg.2010.129. 99, 127, 128, 131, 132

[255] S. Petrović, M. Osborne, and V. Lavrenko. Streaming first story detection with application to Twitter. In Human Language Technologies: the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 181–189, 2010. 99

[256] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In Proc. of the 3rd International Conference on Web Search and Web Data Mining, pages 291–300, 2010. DOI: 10.1145/1718487.1718524.

[257] Hila Becker, Mor Naaman, and Luis Gravano. Beyond trending topics: Real-world event identification on twitter. In Proc. of the 5th International Conference on Weblogs and Social Media (ICWSM), 2011. 99

[258] H. Sayyadi, M. Hurst, and A. Maykov. Event detection and tracking in social streams. In Proc. of the 3rd International ICWSM Conference, pages 311–314, 2009. 99

[259] Takeshi Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In Proc. of the 19th International Conference on World Wide Web (WWW),

pages 851–860. ACM, 2010. DOI: 10.1145/1772690.1772777. 99

[260] J. Y. Weng, C. L. Yang, B. N. Chen, Y. K. Wang, and S. D. Lin. IMASS: An intelligent microblog analysis and summarization system. In Proc. of the ACL-HLT System Demonstrations, pages 133–138, Portland, Oregon, 2011. 99, 127, 131

[261] M. Nagarajan, K. Gomadam, A. Sheth, A. Ranabahu, R. Mutharaju, and A. Jadhav. Spatio-temporal-thematic analysis of citizen sensor data: Challenges and experiences. In Web Information Systems Engineering, pages 539–553, 2009. DOI: 10.1007/978-3-642- 04409-0_52. 100, 127, 128, 129

[262] A. Marcus, M. S. Bernstein, O. Badar, D. R. Karger, S. Madden, and R. C. Miller. TwitInfo: Aggregating and visualizing microblogs for event exploration. In Proc. of the Conference on Human Factors in Computing Systems (CHI), pages 227–236, 2011. DOI: 10.1145/1978942.1978975. 100, 127, 128, 131

[263] M. Naaman, J. Boase, and C. Lai. Is it really about me? Message content in social awareness streams. In Proc. of the ACM Conference on Computer Supported Cooperative Work, pages 189–192. ACM, 2010. DOI: 10.1145/1718918.1718953. 100, 102, 124, 125, 126

[264] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. JASIST, 60(11), pages 2169–2188, 2009. DOI: 10.1002/asi.21149. 100

[265] Patrick Lai. Extracting strong sentiment trends from twitter. http://nlp.stanford.e du/courses/cs224n/2011/reports/patlai.pdf, 2010. 100

[266] Diana Maynard, Kalina Bontcheva, and Dominic Rout. Challenges in developing opinion mining tools for social media. In Proc. of @NLP can u tag #usergeneratedcontent?! Workshop at LREC 2012, Turkey, 2012. 100

[267] A. Pak and P. Paroubek. Twitter based system: Using twitter for disambiguating sentiment ambiguous adjectives. In Proc. of the 5th International Workshop on Semantic Evaluation, pages 436–439, 2010. 101

[268] A. Go, R. Bhayani, and L. Huang. Twitter sentiment classification using distant supervision. Technical Report CS224N Project Report, Stanford University, 2009. 101

[269] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In Proc. of the IEEE Conference on Visual Analytics Science and Technology, pages 115–122, 2010. DOI: 10.1109/vast.2010.5652922. 101, 127

[270] D. Maynard and A. Funk. Automatic detection of political opinions in tweets. In Dieter Fensel Raúl García-Castro and Grigoris Antoniou, Eds., the Semantic Web: ESWC 2011 Selected Workshop Papers, Lecture Notes in Computer Science. Springer, 2011. 101

[271] D. Maynard, M. A. Greenwood, I. Roberts, G. Windsor, and K. Bontcheva. Real-time social media analytics through semantic annotation and linked open data. In Proc. of Web-Sci, Oxford, UK, 2015. DOI: 10.1145/2786451.2786500. 101

[272] M. Dowman, V. Tablan, H. Cunningham, and B. Popov. Web-assisted annotation, semantic indexing and search of television and radio news. In Proc. of the 14th International World Wide Web Conference, Chiba, Japan, 2005. DOI: 10.1145/1060745.1060781. 102

[273] A. Hubmann-Haidvogel, A. M. P. Brasoveanu, A. Scharl, M. Sabou, and S. Gindl. Visualizing contextual and dynamic features of micropost streams. In Proc. of the #MSM2012 Workshop, CEUR, volume 838, 2012. 102, 127, 128, 131

[274] W. X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H. Yan, and X. Li. Comparing twitter and traditional media using topic models. In Proc. of the 33rd European Conference on Advances in Information Retrieval (ECIR), pages 338–349, 2011. DOI: 10.1007/978-3-642-20161- 5_34. 102

[275] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. Misinformation and its correction continued influence and successful debiasing. Psychological

Science in the Public Interest, 13(3), pages 106–131, 2012. DOI: 10.1177/1529100612451018. 102

[276] Rob Procter, Jeremy Crump, Susanne Karstedt, Alex Voss, and Marta Cantijoch. Reading the riots: What were the police doing on twitter? Policing and Society, 23(4), pages 413– 436, 2013. DOI: 10.1080/10439463.2013.780223. 102

[277] Mendoza Marcelo, Poblete Barbara, and Castillo Carlos. Twitter under crisis: Can we trust what we are? In 1st Workshop on Social Media Analytics (SOMA), 2010. DOI: 10.1145/1964858.1964869. 102

[278] Rob Procter, Farida Vis, and Alex Voss. Reading the riots on twitter: Methodological innovation for the analysis of big data. International Journal of Social Research Methodology, 16(3), pages 197–214, 2013. DOI: 10.1080/13645579.2013.774172. 102

[279] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. Early detection of rumors in social media from enquiry posts. In International World Wide Web Conference Committee (IW3C2), 2015. 102, 103

[280] Jing Ma, Wei Gao, Zhongyu Wei, Yueming Lu, and Kam-Fai Wong. Detect rumors using time series of social context information on microblogging websites. In Proc. of the 24th ACM International on Conference on Information and Knowledge Management, CIKM'15, pages 1751–1754, New York, NY, 2015. ACM. 102

[281] Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. Rumor has it: Identifying misinformation in microblogs. In Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNLP'11, pages 1589–1599, 2011. 103

[282] Sardar Hamidian and Mona T Diab. Rumor identification and belief investigation on twitter. In Proc. of NAACL-HLT, pages 3–8, 2016. DOI: 10.18653/v1/w16-0403. 103

[283] Michal Lukasik, Kalina Bontcheva, Trevor Cohn, Arkaitz Zubiaga, Maria Liakata, and Rob Procter. Using Gaussian processes for rumour stance classification in social media. CoRR, abs/1609.01962, 2016. 103

[284] Xiaomo Liu, Armineh Nourbakhsh, Quanzhi Li, Rui Fang, and Sameena Shah. Real- time rumor debunking on twitter. In Proc. of the 24th ACM International on Conference on Information and Knowledge Management, CIKM'15, pages 1867–1870, New York, NY, 2015. ACM. 103

[285] Michal Lukasik, Trevor Cohn, and Kalina Bontcheva. Classifying tweet level judgements of rumours in social media. In Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNLP Lisbon, Portugal, pages 2590–2595, 2015. DOI: 10.18653/v1/d15-1311. 103

[286] Li Zeng, Kate Starbird, and Emma S. Spiro. # unconfirmed: Classifying rumor stance in crisis-related social media messages. In 10th International AAAI Conference on Web and Social Media, 2016. 103

[287] Arkaitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. Analysing how people orient to and spread rumours in social media by looking at conversational threads. PLoS ONE, 11(3), pages 1–29, 2016. DOI: 10.1371/jour- nal.pone.0150989. 103

[288] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. ZenCrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In Proc. of the 21st Conference on World Wide Web, pages 469–478, 2012. DOI: 10.1145/2187836.2187900. 103

[289] Christian M. Meyer and Iryna Gurevych. Wiktionary: A new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography. In Electronic Lexicography. Oxford University Press, 2012. DOI: 10.1093/acprof:oso/9780199654864.003.0013. 104, 136

[290] Iryna Gurevych, Judith Eckle-Kohler, Silvana Hartmann, Michael Matuschek, Chris- tian M. Meyer, and Christian Wirth. UBY: A large-scale unified lexical-semantic resource based on LMF. In 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 580–590, 2012. 104, 136

[291] Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. Towards linguistically grounded ontologies. In Proc. of the European Semantic Web Conference (ESWC'09), LNCS 5554, pages 111—125, 2009. DOI: 10.1007/978-3-642-02121-3_12. 104, 136

[292] T. Finin, W. Murnane, A. Karandikar, N. Keller, J. Martineau, and M. Dredze. Annotating named entities in twitter data with crowdsourcing. In Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 80–88, 2010. 104, 139

[293] D. Maynard and M. A. Greenwood. Large scale semantic annotation, indexing and search at the national archives. In Proc. of LREC 2012, Turkey, 2012. 105, 113, 138

[294] Kalina Bontcheva, Valentin Tablan, and Hamish Cunningham. Semantic search over documents and ontologies. In Bridging Between Information Retrieval and Databases, volume 8173, pages 31–53. Springer Verlag, 2014. DOI: 10.1007/978-3-642-54798-0_2. 107

[295] V. Tablan, K. Bontcheva, I. Roberts, and H. Cunningham. Mímir: An open-source semantic search framework for interactive information seeking and discovery. Journal of Web Semantics, 30, pages 52–68, 2015. DOI: 10.1016/j.websem.2014.10.002. 107, 110, 111, 113, 120

[296] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, and M. Goranov. Semantic annotation, indexing and retrieval. Journal of Web Semantics, 1(2), pages 671–680, 2004. DOI: 10.1016/j.websem.2004.07.005. 107, 109, 112

[297] Hamish Cunningham, Valentin Tablan, Ian Roberts, Mark A. Greenwood, and Niraj Aswani. Information extraction and semantic annotation for multi-paradigm information management. In Mihai Lupu, Katja Mayer, John Tait, and Anthony J. Trippe, Eds., Current Challenges in Patent Information Retrieval, volume 29 of the Information Retrieval Series, pages 307–327. Springer Berlin Heidelberg, 2011. DOI: 10.1007/978-3-642-19231- 9. 107, 109, 138

[298] K. Mahesh, J. Kud, and P. Dixon. Oracle at TREC8: A lexical approach. In Proc. of the 8th Text Retrieval Conference (TREC-8), 1999. 107

[299] E. Voorhees. Using WordNet for text retrieval. In C. Fellbaum, Ed., WordNet: An Electronic Lexical Database. MIT Press, 1998. 107

[300] Li Ding, Tim Finin, Anupam Joshi, Rong Pan, R. Scott Cost, Yun Peng, Pavan Reddivari, Vishal C. Doshi, and Joel Sachs. Swoogle: A search and metadata engine for the semantic web. In Proc. of the 13th ACM Conference on Information and Knowledge Management, 2004. DOI: 10.1145/1031171.1031289. 108

[301] M. Hildebrand, J. van Ossenbruggen, and J. Hardman. Facet: A browser for heterogeneous semantic web repositories. In Proc. of the 5th International Semantic Web Conference, 2006. DOI: 10.1007/11926078_20. 108

[302] G. Klyne and J. Carroll. Resource description framework (RDF): Concepts and abstract syntax. W3C recommendation, W3C, 2004. http://www.w3.org/TR/rdf-concepts/ 108

[303] Sean Bechhofer, Frank van Harmelen, Jim Hendler, Ian Horrocks, Deborah L. McGuinness, Peter F. Patel-Schneider, and Lynn Andrea Stein. OWL web ontology language reference. W3C recommendation, W3C, http://www.w3.org/, 2004. 108

[304] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query language for RDF. W3C recommendation, W3C, http://www.w3.org/TR/rdf-sparql-query/, 2008. 108

[305] Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haussmann. A case for semantic full-text search. In Proc. of the 1st Joint International Workshop on Entity-Oriented and Semantic Search, JIWES'12, pages 4:1–4:3. ACM, 2012. DOI: 10.1145/2379307.2379311. 109

[306] Hannah Bast, Florian Bäurle, Björn Buchhold, and Elmar Haussmann. Broccoli: Semantic full-text search at your fingertips. CoRR, abs/1207.2615, 2012. 109, 110, 111

[307] Amit Singhal. Introducing the knowledge graph: things, not strings. http://google blog.blogspot.it/2012/05/introducing-knowledge-graph-things-not.html, 2012. 109

[308] K. Bontcheva, J. Kieniewicz, S. Andrews, and M. Wallis. Semantic enrichment and search: A case study on environmental science literature. D-Lib Magazine, 21(1/2), 2015. DOI: 10.1045/january2015-bontcheva. 110, 116

[309] J. Kieniewicz, A. Sudlow, and E. Newbold. Coordinating improved environmental information access and discovery: Innovations in sharing environmental observations and information. In W. Pillman, S. Schade, and P.Smits, Eds., Proc. of the 25th International EnviroInfo Conference, 2011. 110

[310] J. Kieniewicz and M. Wallis. User requirements. Technical Report http://gate.ac.uk/projects/envilod/EnviLOD-WP2-User-Requirements.pdf, EnviLOD project deliverable, 2012. 110

[311] Mihai Lupu and Allan Hanbury. Patent retrieval. Foundations and Trends in Information Retrieval, 7(1), pages 1–97, 2013. DOI: 10.1561/1500000027. 110

[312] Lei Zhang, Qiaoling Liu, Jie Zhang, Haofen Wang, Yue Pan, and Yong Yu. Semplore: An ir approach to scalable hybrid query of semantic web data. In the Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC-ASWC, pages 652–665, 2007. DOI: 10.1007/978-3-540-76298-0_47. 110, 111

[313] Miriam Fernández, Iván Cantador, Vanesa López, David Vallet, Pablo Castells, and Enrico Motta. Semantically enhanced information retrieval: An ontology-based approach. Web Semantics, 9(4), pages 434–452, 2011. DOI: 10.1016/j.websem.2010.11.003. 111

[314] Haofen Wang, thanh Tran, Chang Liu, and Linyun Fu. Lightweight integration of ir and db for scalable hybrid search with integrated ranking support. Web Semantics: Science, Services and Agents on the World Wide Web, 9(4), 2011. DOI: 10.1016/j.websem.2011.08.002. 111

[315] Bettina Fazzinga, Giorgio Gianforme, Georg Gottlob, and thomas Lukasiewicz. Semantic web search based on ontological conjunctive queries. Web Semantics: Science, Services and Agents on the World Wide Web, 9(4), 2011. DOI: 10.1016/j.websem.2011.08.003. 111

[316] Nikos Bikakis, Giorgos Giannopoulos, Theodore Dalamagas, and Timos Sellis. Integrating keywords and semantics on document annotation and search. In Robert Meersman, tharam Dillon, and Pilar Herrero, Eds., On the Move to Meaningful Internet Systems, volume 6427, pages 921–938. Springer, 2010. DOI: 10.1007/978-3-540-88871-0. 111

[317] Borislav Popov, Atanas Kiryakov, Angel Kirilov, Dimitar Manov, Damyan Ognyanoff, and Miroslav Goranov. KIM—Semantic annotation platform. In 2nd International Semantic Web Conference (ISWC), pages 484–499, Berlin, 2003. Springer. DOI: 10.1007/978-3-540-39718-2_53. 112, 132

[318] Atanas Kiryakov. OWLIM: Balancing between scalable repository and light-weight reasoner. In Proc. of the 15th International World Wide Web Conference (WWW), 2010, Edinburgh, Scotland, 2006. 113

[319] Paolo Boldi and Sebastiano Vigna. MG4J at TREC 2005. In Ellen M. Voorhees and Lori P. Buckland, Eds., Proc. of the 14th Text REtrieval Conference (TREC), volume 500 of Special Publications, pages 266–271. NIST, 2005. http://mg4j.dsi.unimi.it/ 113

[320] V. Tablan, I. Roberts, H. Cunningham, and K. Bontcheva. Gatecloud.net: A platform for large-scale, open-source text processing on the cloud. Philosophical Transactions of the Royal Society A, 371(1983), 2013. DOI: 10.1098/rsta.2012.0071. 113, 138

[321] J. Teevan, D. Ramage, and M. R. Morris. #twittersearch: A comparison of microblog search and web search. In Proc. of the 4th ACM International Conference on Web Search and Data Mining (WSDM), pages 35–44, 2011. DOI: 10.1145/1935826.1935842. 118

[322] K. Bontcheva and D. Rout. Making sense of social media through semantics: A survey. Semantic Web—Interoperability, Usability, Applicability, 5(5), pages 373—403, 2014. 119

[323] K. Holmberg and I. Hellsten. Analyzing the climate change debate on twitter—content and differences between genders. In Proc. of the ACM WebScience Conference, pages 287– 288, Bloomington, IN, 2014. DOI: 10.1145/2615569.2615638.

[324] René Pfitzner, Antonios Garas, and Frank Schweitzer. Emotional divergence influences information spreading in twitter. ICWSM, 12, pages 2–5, 2012.

[325] C. Meili, R. Hess, M. Fernandez, and G. Burel. Earth hour report. Technical Report D6.2.1, DecarboNet Project Deliverable, 2014.

[326] Matthew Rowe and Harith Alani. Mining and comparing engagement dynamics across multiple social media platforms. In Proc. of the ACM conference on Web science, pages 229– 238, 2014. DOI: 10.1145/2615569.2615677. 119

[327] A. Esuli and F. Sebastiani. SentiWordNet: A publicly available lexical resource for opinion mining. In Proc. of LREC, 2006. 119

[328] Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. GATE: An architecture for development of robust HLT applications. In Proc. of the 40th Annual Meeting on Association for Computational Linguistics, ACL'02, pages 168–175, Stroudsburg, PA, 2002. Association for Computational Linguistics. DOI: 10.3115/1073083.1073112. 119

[329] K. Tao, F. Abel, C. Hauff, and G.-J. Houben. What makes a tweet relevant to a topic. In Proc. of the #MSM2012 Workshop, CEUR, volume 838, 2012. 119

[330] P. N. Mendes, A. Passant, and P. Kapanipathi. Twarql: Tapping into the wisdom of the crowd. In Proc. of the 6th International Conference on Semantic Systems, I-SEMANTICS'10, pages 45:1–45:3, 2010. DOI: 10.1145/1839707.1839762. 119

[331] F. Abel, I. Celik, G.-J. Houben, and P. Siehndel. Leveraging the semantics of tweets for adaptive faceted search on Twitter. In Proc. of the 10th International Conference on the Semantic Web—Volume Part I, ISWC'11, pages 1–17, Berlin, Heidelberg, 2011. Springer-Verlag. DOI: 10.1007/978-3-642-25073-6_1. 120

[332] Miriam Fernandez, Arno Scharl, Kalina Bontcheva, and Harith Alani. User profile modelling in online communities. In Proc. of the 3rd International Conference on Semantic Web Collaborative Spaces—Volume 1275, pages 1–15. CEUR-WS. org, 2014. 122

[333] L. Aroyo and G.-J. Houben. User modeling and adaptive semantic web. Semantic Web, 1(1, 2), pages 105–110, 2010. DOI: 10.3233/SW-2010-0006. 122

[334] S. Decker and M. Frank. the Social Semantic Desktop. Technical report, DERI Technical Report 2004-05-02, 2004. 122

[335] P. Mika. Ontologies are us: A unified model of social networks and semantics. Journal of Web Semantics, 5(1), pages 5–15, 2007. DOI: 10.1007/11574620_38. 122

[336] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: Experiments on recommending content from information streams. In Proc. of the 28th International Conference on Human Factors in Computing Systems, CHI'10, pages 1185–1194, 2010. DOI: 10.1145/1753326.1753503. 123, 126

[337] P. Kapanipathi, F. Orlandi, A. Sheth, and A. Passant. Personalized filtering of the twitter stream. In 2nd workshop on Semantic Personalized Information Management at ISWC, 2011. 123, 124

[338] M. Szomszor, H. Alani, I. Cantador, K. O'Hara, and N. Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis. In Proc. of the 7th International Conference on the Semantic Web (ISWC), pages 632–648. Springer-Verlag, 2008. DOI: 10.1007/978-3-540-88564-1_40. 123

[339] E. Zavitsanos, G. A. Vouros, and G. Paliouras. Classifying users and identifying user interests in folksonomies. In Proc. of the 2nd Workshop on Semantic Personalized Information Management: Retrieval and Recommendation, 2011. 123

[340] M. Zhou, S. Bao, X. Wu, and Y. Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In Proc. of the 6th International Semantic Web Conference, ISWC'07, pages 680–693. Springer-Verlag, 2007. DOI: 10.1007/978-3-540-76298-0_49. 123

[341] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In Proc. of the 7th International Conference on the Semantic Web, pages 615–631, 2008. DOI: 10.1007/978-3-540-88564-1_39. 123

[342] Zhiyuan Cheng, James Caverlee, and Kyumin Lee. You are where you tweet: A content- based approach to geo-locating twitter users. In Proc. of the 19th ACM International Conference on Information and Knowledge Management, CIKM'10, pages 759–768, New York, NY, 2010. ACM. DOI: 10.1145/1871437.1871535. 123

[343] S. Yardi and D. Boyd. Tweeting from the town square: Measuring geographic local networks. In Proc. of ICWSM, 2010. 123

[344] J. Burger, J. Henderson, G. Kim, and G. Zarrella. Discriminating Gender on Twitter. In Proc. of the Conference on Empirical Methods in Natural Language Processing, EMNLP'11, pages 1301–1309, 2011. 123

[345] M. Pennacchiotti and A. M. Popescu. A machine learning approach to twitter user classification. In Proc. of ICWSM, pages 281–288, 2011. 123

[346] Clay Fink, Christine Piatko, James Mayfield, Tim Finin, and Justin Martineau. Geolocating blogs from their textual content. In Working Notes of the AAAI Spring Symposium on Social Semantic Web: Where Web 2.0 Meets Web 3.0, pages 1–2. AAAI Press, 2008. 123

[347] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 1277–1287, 2010. 123

[348] Jalal Mahmud, Jeffrey Nichols, and Clemens Drews. Where is this tweet from? Inferring home locations of twitter users. In Proc. of the 6th International AAAI Conference on Weblogs and Social Media, Dublin, Ireland, 2012. 123

[349] J. Chan, C. Hayes, and E. Daly. Decomposing discussion forums using common user roles. In Proc. of WebSci10: Extending the Frontiers of Society On-Line, 2010. 124

[350] Markus Strohmaier, Christian Koerner, and Roman Kern. Why do users tag? detecting users' motivation for tagging in social tagging systems. 2010. 124

[351] A. L. Gentile, V. Lanfranchi, S. Mazumdar, and F. Ciravegna. Extracting semantic user networks from informal communication exchanges. In Proc. of the 10th International Conference on the Semantic Web, ISWC'11, pages 209–224. Springer-Verlag, 2011. DOI: 10.1007/978-3-642-25073-6_14. 125

[352] C. Beaudoin. Explaining the relationship between internet use and interpersonal trust: Taking into account motivation and information overload. Journal of Computer Mediated Communication, 13, pages 550—568, 2008. DOI: 10.1111/j.1083-6101.2008.00410.x. 125

[353] Fabian Abel, Qi Gao, Geert-Jan Houben, and Ke Tao. Analyzing temporal dynamics in twitter profiles for personalized recommendations in the social web. In Proc. of the 3rd International Web Science Conference, WebSci'11, pages 2:1–2:8, New York, NY, 2011. ACM. DOI: 10.1145/2527031.2527040. 126

[354] J. Chen, R. Nairn, and E. Chi. Speak little and well: Recommending conversations in online social streams. In Proc. of the Annual Conference on Human Factors in Computing Systems, CHI'11, pages 217–226, 2011. DOI: 10.1145/1978942.1978974. 126, 132

[355] N. Bansal and N. Koudas. Blogscope: Spatio-temporal analysis of the blogosphere. In Proc. of the 16th International Conference on World Wide Web, WWW'07, pages 1269–1270, 2007. DOI: 10.1145/1242572.1242802. 127, 128, 132

[356] D. A. Shamma, L. Kennedy, and E. F. Churchill. Tweetgeist: Can the twitter timeline reveal the structure of broadcast events? In Proc. of CSCW, 2010. 127, 131

[357] O. Phelan, K. McCarthy, and B. Smyth. Using twitter to recommend real-time topical news. In Proc. of the ACM Conference on Recommender Systems, pages 385–388, 2009. DOI: 10.1145/1639714.1639794. 127

[358] M. S. Bernstein, B. Suh, L. Hong, J. Chen, S. Kairam, and E. H. Chi. EDDI: Interactive topic-based browsing of social status streams. In Proc. of the 23nd ACM Symposium on User Interface Software and Technology (UIST), pages 303–312, 2010. DOI: 10.1145/1866029.1866077. 127

[359] B. Adams, D. Phung, and S. Venkatesh. Eventscapes: Visualizing events over time with emotive facets. In Proc. of the 19th ACM International Conference on Multimedia, pages 1477–1480, 2011. DOI: 10.1145/2072298.2072044. 127, 131

[360] J. Eisenstein, D. H. P. Chau, A. Kittur, and E. Xing. Topicviz: Semantic navigation of document collections. In CHI Work-in-Progress Paper (Supplemental Proceedings), 2012. 128

[361] D. Archambault, D. Greene, P. Cunningham, and N. J. Hurley. themeCrowds: Multiresolution summaries of twitter usage. In Workshop on Search and Mining User-generated Contents (SMUC), pages 77–84, 2011. DOI: 10.1145/2065023.2065041. 128, 132

[362] B. Meyer, K. Bryan, Y. Santos, and B. Kim. TwitterReporter: Breaking news detection and visualization through the geo-tagged twitter network. In Proc. of the ISCA 26th International Conference on Computers and their Applications, pages 84–89, 2011. 128, 131

[363] S. Faridani, E. Bitton, K. Ryokai, and K. Goldberg. Opinion space: A scalable tool for browsing online comments. In Proc. of

the 28th International Conference on Human Factors in Computing Systems (CHI), pages 1175–1184, 2010. DOI: 10.1145/1753326.1753502. 131

[364] Omar F. Zaidan and Chris Callison-Burch. Crowdsourcing translation: Professional quality from non-professionals. In Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 1220–1229, 2011. 136

[365] Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. Using mechanical turk to create a corpus of Arabic summaries. In Proc. of the 7th Conference on International Language Resources and Evaluation, 2010. 136

[366] Vamshi Ambati and Stephan Vogel. Can crowds build parallel corpora for machine translation systems? In Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 62–65, 2010. 136

[367] Chris Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In Proc. of the Conference on Empirical Methods in Natural Language Processing, pages 286–295, 2009. DOI: 10.3115/1699510.1699548. 139

[368] Ann Irvine and Alexandre Klementiev. Using mechanical turk to annotate lexicons for less commonly used languages. In Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 108–113, 2010. 136

[369] A. Weichselbraun, S. Gindl, and A. Scharl. Using games with a purpose and bootstrapping to create domain-specific sentiment lexicons. In Proc. of the 20th ACM Conference on Information and Knowledge Management (CIKM), pages 1053–1060, 2011. DOI: 10.1145/2063576.2063729. 136

[370] Nitin Madnani, Jordan Boyd-Graber, and Philip Resnik. Measuring transitivity using untrained annotators. In Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 188–194, 2010. 136

[371] Massimo Poesio, Nils Diewald, Maik Stührenberg, Jon Chamberlain, Daniel Jettka, Daniela Goecke, and Udo Kruschwitz. Markup infrastructure for the anaphoric bank: Supporting web collaboration. In Alexander Mehler, Kai-Uwe Kühnberger, Henning Lobin, Harald Lüngen, Angelika Storrer, and Andreas Witt, Eds., Modeling, Learning, and Processing of Text Technological Data Structures, volume 370 of Studies in Computational Intelligence, pages 175–195. Springer Berlin/Heidelberg, 2012. DOI: 10.1007/978-3-642- 22613-7. 136, 138, 139

[372] W. Rafelsberger and A. Scharl. Games with a purpose for social networking platforms. In Proc. of the 20th ACM Conference on Hypertext and hypermedia, HT'09, pages 193–198, 2009. DOI: 10.1145/1557914.1557948. 136

[373] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D. Manning. Multi-instance multi-label learning for relation extraction. In Jun'ichi Tsujii, James Henderson, and Marius Pasça, Eds., Proc. of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pages 455–465, Jeju Island, Korea, 2012. Association for Computational Linguistics. 137

[374] Vidhoon Viswanathan, Nazneen Fatema Rajani, Yinon Bentor, and Raymond Mooney. Stacked ensembles of information extractors for knowledge-base population. In Chengqing Zong and Michael Strube, Eds., Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 177–187, Beijing, China, 2015. Association for Computational Linguistics. DOI: 10.3115/v1/p15-1. 137

[375] Alon Halevy, Peter Norvig, and Fernando Pereira. the unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2), pages 8–12, 2009. DOI: 10.1109/mis.2009.36. 137

[376] Roger Barga, Dennis Gannon, and Daniel Reed. the client and the cloud: Democratizing research computing. IEEE Internet Computing, 15(1), pages 72–75, 2011. DOI: 10.1109/mic.2011.20. 137

[377] Marios D. Dikaiakos, Dimitrios Katsaros, Pankaj Mehra, George Pallis, and Athena Vakali. Cloud computing: Distributed internet computing for IT and scientific research. IEEE Internet Computing, 13(5), pages 10–13, 2009. DOI: 10.1109/mic.2009.103. 137

[378] Kalina Bontcheva, Hamish Cunningham, Ian Roberts, Angus Roberts, Valentin Tablan, Niraj Aswani, and Genevieve Gorrell. GATE Teamware: A web-based, collaborative text annotation framework. Language Resources and Evaluation, 47, pages 1007—1029, 2013. DOI: 10.1007/s10579-013-9215-6. 139

[379] Seid Muhie Yimam, Chris Biemann, Richard Eckart de Castilho, and Iryna Gurevych. Automatic annotation suggestions and custom annotation layers in webanno. In Proc. of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 91–96, Baltimore, Maryland, 2014. Association for Computational Linguistics. DOI: 10.3115/v1/p14-5016. 139

[380] Leah Hoffmann. Crowd control. Communications of the ACM, 52(3), pages 16–17, 2009. DOI: 10.1145/1467247.1467254. 139

[381] Sharoda A. Paul, Lichan Hong, and Ed H. Chi. What is a question? Crowdsourcing tweet categorization. In CHI'2011 Workshop on Crowdsourcing and Human Computation, 2011. 139

[382] K. Siorpaes and M. Hepp. Games with a purpose for the semantic web. Intelligent Systems, IEEE, 23(3), pages 50–60, 2008. DOI: 10.1109/mis.2008.45. 139

[383] S. thaler, K. S. E. Simperl, and C. Hofer. A survey on games for knowledge acquisition. Technical Report Tech. Rep. STI TR 2011-05-01, Semantic Technology Institute, 2011. 139

[384] J. Waitelonis, N. Ludwig, M. Knuth, and H. Sack. WhoKnows? Evaluating linked data heuristics with a quiz that cleans up DBpedia. Interactive Technology and Smart Education, 8(4), pages 236–248, 2011. DOI: 10.1108/17415651111189478. 139

[385] Richard McCreadie, Craig Macdonald, and Iadh Ounis. Identifying top news using crowdsourcing. Information Retrieval, pages 1–31, 2012. 10.1007/s10791-012-9186-z. DOI: 10.1007/s10791-

012-9186-z. 139

[386] Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In Proc. of the NAACL HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 148–151, 2010. 139

[387] David Inouye and Jugal K. Kalita. Comparing twitter summarization algorithms for multiple post summaries. In SocialCom/PASSAT, pages 298–306, 2011. DOI: 10.1109/pas- sat/ socialcom.2011.31. 139

[388] Andrea Glaser and Hinrich Schütze. Automatic generation of short informative sentiment summaries. In Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics, pages 276–285, Avignon, France, 2012. 139

[389] Kalina Bontcheva, Ian Roberts, Leon Derczynski, and Dominic Rout. the GATE crowdsourcing plugin: Crowdsourcing annotated corpora made easy. In Proc. of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL). Association for Computational Linguistics, 2014. DOI: 10.3115/v1/ e14-2025. 139

[390] G.W. Allport and L. Postman. the psychology of rumor. Journal of Clinical Psychology, 1947. 102

دراسات ١٦

معالجة اللغات الطبيعية للويب الدلالي

يعمل مركز الملك عبدالله بن عبدالعزيز الدولي لخدمة اللغة العربية على تعزيز خدماته في المجالات المتنوعة لخدمة اللغة العربية وعلومها، إذ ينطلق من رؤية موحّدة في أعماله عامة - ومنها برنامج النشر - وذلك بأن يطلق برامجه ودراساته في المجالات التي تفتقر إلى جهود نوعية، أو التي تحتاج إلى تكثيف العمل فيها.

ويجتهد المركز في انتقاء الكتب التي تصدر ضمن هذه السلسلة، بأن تكون مضيفة إلى حقلها المعرفي، ومفتاحا للمشروعات العلمية والعملية، ومحققة لتراكمٍ معرفيٍّ مثرٍ.

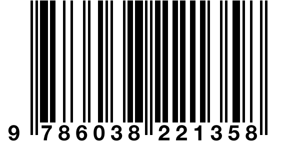
وإذ تشيد الأمانة العامة بجهد مترجم الكتاب، ترجمةً، وتصحيحًا لمسوداته، ومراجعةً للطباعة، فإنها تدعو الباحثين كافة من أنحاء العالم إلى المساهمة في هذه السلسة، لتتكامل مع سلاسل المركز العلمية الأخرى.

ويسعد المركز بالعمل مع المؤسسات والأفراد المختصين والمهتمين في خدمة لغتنا العربية، وتكثيف الجهود والتكامل نحو تمكين لغتنا، وتحقيق وجودها السامي في مجالات الحياة.



ص.ب ١٢٥٠٠ الرياض ١١٤٧٣
هاتف:٠٠٩٦٦١١٢٥٨٧٢٦٨ - ٠٠٩٦٦١١٢٥٨١٠٨٢
البريد الإليكتروني: nashr@kaica.org.sa




9 786038 221358